Northumbria University
NEWCASTLE

RE-IDENTIFYING PEOPLE IN THE CROWD

CHIRINE RIACHY

PhD

2019

RE-IDENTIFYING PEOPLE IN THE CROWD

CHIRINE RIACHY

A thesis submitted in partial fulfilment of the requirements of the University of Northumbria at Newcastle for the degree of Doctor of Philosophy

Faculty of Engineering and Environment

September 2019

To my mother, Samia

Abstract

Developing an automated surveillance system is of great interest for various reasons including forensic and security applications. In the case of a network of surveillance cameras with non-overlapping fields of view, person detection and tracking alone are insufficient to track a subject of interest across the network. In this case, instances of a person captured in one camera view need to be retrieved among a gallery of different people, in other camera views. This vision problem is commonly known as person re-identification (re-id).

Cross-view instances of pedestrians exhibit varied levels of illumination, viewpoint, and pose variations which makes the problem very challenging. Despite recent progress towards improving accuracy, existing systems suffer from low applicability to real-world scenarios. This is mainly caused by the need for large amounts of annotated data from pairwise camera views to be available for training. Given the difficulty of obtaining such data and annotating it, this thesis aims to bring the person re-id problem a step closer to real-world deployment.

In the first contribution, the single-shot protocol, where each individual is represented by a pair of images that need to be matched, is considered. Following the extensive annotation of four datasets for six attributes, an evaluation of the most widely used feature extraction schemes is conducted. The results reveal two high-performing descriptors among those evaluated, and show illumination variation to have the most impact on re-id accuracy.

Motivated by the wide availability of videos from surveillance cameras and the additional visual and temporal information they provide, video-based person re-id is then investigated, and a supervised system is developed. This is achieved by improving and extending the best performing image-based person descriptor into three dimensions and combining it with distance metric learning. The system obtained achieves state-of-the-art results on two widely used datasets.

Given the cost and difficulty of obtaining labelled data from pairwise cameras in a network to train the model, an unsupervised video-based person re-id method is also developed. It is based on a set-based distance measure that leverages rank vectors to estimate the similarity scores between person tracklets. The proposed system outperforms other unsupervised methods by a large margin on two datasets while competing with deep learning methods on another large-scale dataset.

List of Publications

Published peer-reviewed journal papers:

- C. Riachy, F. Khelifi, and A. Bouridane, "Video-based person re-identification using unsupervised tracklet matching," *IEEE Access*, vol. 7, pp. 20596–20606, 2019.
- D. Organisciak, C. Riachy, N. Aslam, and H. Shum, "Triplet loss with channel attention for person re-identification," *Journal of WSCG*, vol. 27(2), pp. 161-169, 2019.

Published peer-reviewed conference papers:

- C. Riachy and A. Bouridane, "Person re-identification: Attribute-based feature evaluation," in *IEEE World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2018.
- C. Riachy, N. Al-Maadeed, D. Organisciak, F. Khelifi, and A. Bouridane, "3D Gaussian descriptor for video-based person re-identification," in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2019.

Contents

List of Figures

List of Tables

Acronyms

APD	Average Point-wise Distance
BIF	Biologically-Inspired Features
BoW	Bag-of-Words
BTF	Brightness Transfer Function
CCTV	Closed-Circuit Television
CMC	Cumulative Matching Characteristic
CNN	Convolutional Neural Network
DAL	Deep Association Learning
DGM	Dynamic Graph Matching
ELF	Ensemble of Localised Features
FDA	Fisher Discriminant Analysis
FEP	Flow Energy Profile
FOV	Field of View
GMM	Gaussian Mixture Model
GOG	Gaussian of Gaussian
GOG3D	3-Dimensional Gaussian of Gaussian
GPU	Graphics Processing Unit

GRID	UnderGround Re-Identification
HOG	Histogram of Oriented Gradients
HOG3D	Histograms of Oriented 3D Gradients
i-LIDS	Imagery Library of Intelligent Detection Systems
IDE	ID-discriminative Embedding
KISSME	Keep It Simple and Straightforward Metric
kLFDA	Kernel Local Fisher Discriminant Analysis
kMFA	Kernel Marginal Fisher Analysis
kNFST	Kernel Null Foley-Sammon Transform
kPCCA	Kernel Pairwise Constrained Component Analysis
LBP	Local Binary Pattern
LFDA	Local Fisher Discriminant Analysis
LOMO	Local Maximal Occurrence
LPP	Locality Preserving Projections
LSTM	Long Short-Term Memory
MAP	Maximum A Posteriori
mAP	Mean Average Precision
MARS	Motion Analysis and Re-identification Set
MCTS	Multiple-Camera Tracking Scenario
MDTS-DTW	Multi-Dimensional Time Shift Dynamic Time Warping
MFA	Marginal Fisher Analysis
ML	Maximum Likelihood
MPD	Minimum Point-wise Distance
MSCR	Maximally Stable Colour Regions

NBNN	Naive Bayes Nearest Neighbour
NFST	Null Foley-Sammon Transform
PAM	Part-Appearance Mixture
PCA	Principal Component Analysis
PCCA	Pairwise Constrained Component Analysis
PRDC	Probabilistic Relative Distance Comparison
PRID2011	Person Re-ID 2011
PRID450S	Person Re-ID 450S
PSD	Positive Semi-Definite
QAN	Quality Aware Network
RHSP	Recurrent High-Structured Patches
RNN	Recurrent Neural Network
SDALF	Symmetry-Driven Accumulation of Local Features
SDM	Sequential Decision Making
SE	Squeeze and Excitation
SIFT	Scale-Invariant Feature Transform
SIFT3D	3-Dimensional Scale-Invariant Feature Transform
SILTP	Scale-Invariant Local Ternary Pattern
SMP	Stepwise Metric Promotion
SPD	Symmetric Positive Definite
SSS	Small Sample Size
SVM	Support Vector Machine
TAUDL	Tracklet Association Unsupervised Deep Learning
UnKISSME	Unsupervised KISSME

VIPeR	Viewpoint Invariant Pedestrian Recognition
XQDA	Cross-View Quadratic Discriminant Analysis

Acknowledgements

My PhD journey would have been much more difficult if it weren't for the help and support of many people I knew or met along the way.

My deepest gratitude goes to my supervisor, Prof. Ahmed Bouridane, who cheerfully accepted to take over my supervision at a critical stage of my PhD. Thank you for believing in me and allowing me to pursue my ideas while providing priceless support and guidance. Your willingness to help, knowledge, and advice throughout the different stages of my PhD are the reason I made it this far. My gratitude extends to my second supervisor, Dr Fouad Khelifi, whose critical and rigorous feedback helped improve substantially my critical thinking and the quality of my work, and from whom I have learned a lot. I would also like to thank Prof. Ling Shao for initially offering me this position and guiding my first steps in this field.

Special thanks go to Kaveen Perera from the IT department for always finding creative solutions to the many hardware/software issues I encountered. I would also like to thank Daniel Organisciak for fruitful and enjoyable collaborations, endless insightful discussions, and much appreciated moral support.

I am grateful to my examiners, Prof. Farzin Deravi and Dr Neil Eliot, for kindly offering their constructive critique and for providing invaluable insights to improve the quality of this thesis. I especially thank them for making the viva (unexpectedly but happily) a memorable experience.

An immense thank you goes to my office-mates in labs Pandon F7 and Ellison B108 for making the trip to university a rather delightful one and the lab a happy place.

To my friends outside academia, you were the breath of fresh air I most needed. An enormous thank you goes to Geri for putting up with my vents and making me laugh during the most stressful times. Your humour and moral support helped preserve some of my sanity. I would also like to thank my cousins and friends in Lebanon for their constant encouragement.

To my mother, Samia, the only source of unconditional love in this world, you inspire me more every day. I owe you everything I have achieved or ever will. To my sister, Ninar, who looked after the family while I was away and assumed a lot of responsibilities so I can pursue my goals, a big thanks to you and to the most adorable nephews, Mike and Karl, for their understanding. To my

father, Gaby, who encouraged me to take on this step but couldn't see the end of it, I took solace knowing that this would have made you proud. I miss you very much.

Finally, I would like to sincerely acknowledge full financial support from Northumbria University that made this thesis possible.

Declaration

I declare that the work contained in this thesis has not been submitted for any other award and that it is all my own work. I also confirm that this work fully acknowledges opinions, ideas and contributions from the work of others.

Any ethical clearance for the research presented in this thesis has been approved. Approval has been sought and granted by the *Faculty Ethics Committee* on *19/07/2016*.

I declare that the Word Count of this thesis is 37,823 words.

Name: Chirine Riachy

Signature:

Date: 30 September 2019

Chapter 1

Introduction

1.1 Person Re-Identification Problem

The large deployment of surveillance cameras in public places comes alongside a considerable cost if monitoring them was not automated. Having human operators monitor videos captured by a number of cameras in real-time is prohibitively costly. More importantly, it was shown to be highly inefficient due to fatigue and inevitable attentive gaps [1]. Even when real-time monitoring is not needed, manually searching a network of cameras for a specific person or event is immensely impractical. For these reasons, automated visual surveillance has been of great interest lately [2].

A crucial task in automated surveillance across a Closed-Circuit Television (CCTV) network is detecting and tracking people. However, this alone is not sufficient. For practical reasons, the cameras in a CCTV network are usually spread over wide areas and often have non-overlapping Fields of View (FOVs). This requires a subject of interest captured in one camera, to be retrieved (re-identified) in other camera views. In other words, when a person's image or video is available in one camera view (the probe), a set of person images/videos from a different camera view (the gallery) have to be searched in order to match and trace that same person across the network. This problem is known as person re-identification (re-id), and has been a very active research topic recently [3]. The vast majority of the work on re-id did not involve any night vision scenarios until the first benchmark was recently released [4]. Therefore, it is not covered in this thesis. An illustration of the problem can be seen in Figure 1.1.

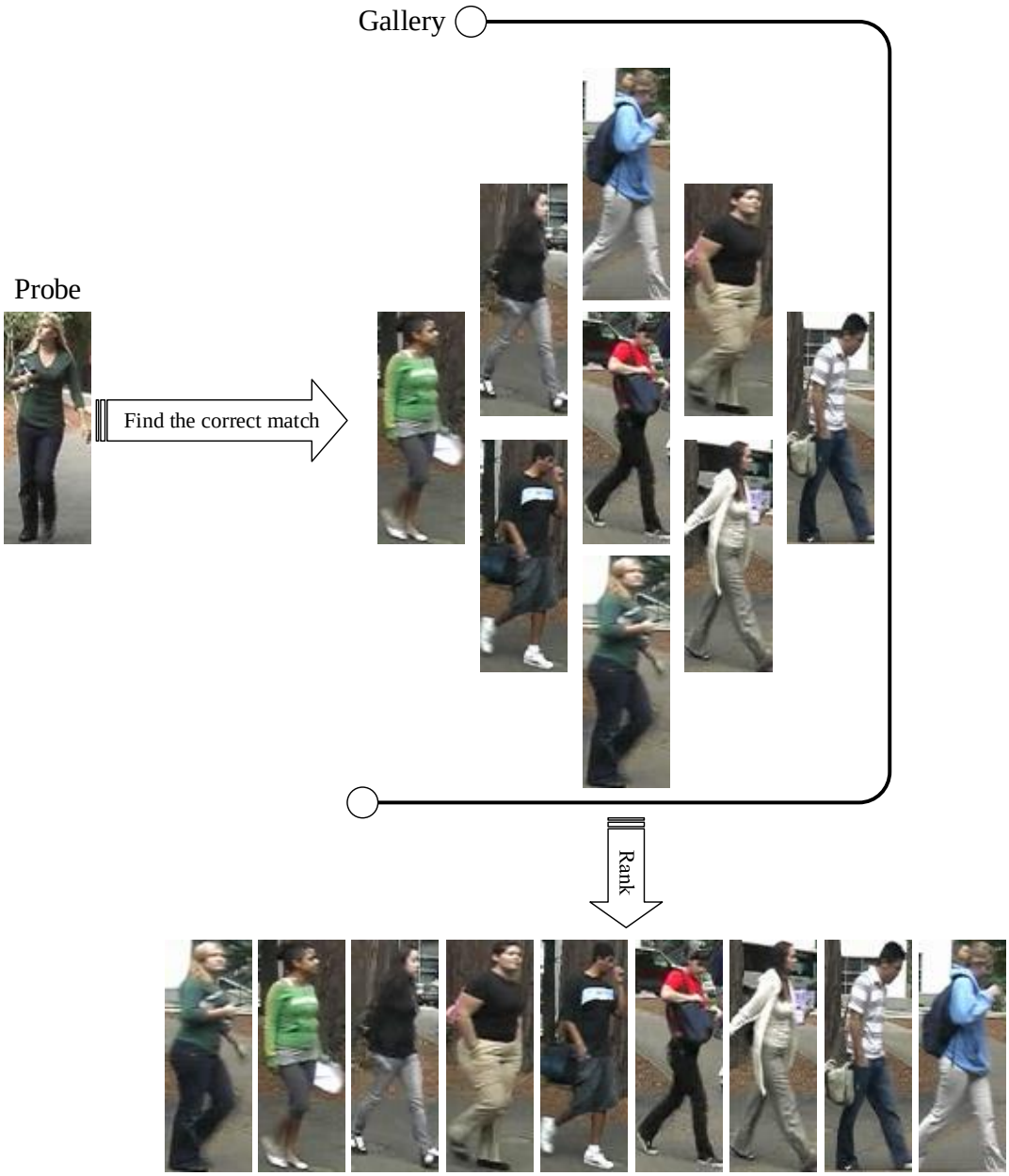


Figure 1.1: The re-identification problem consists of finding instances in the gallery set bearing the same identity as the probe. The gallery elements are ranked according to their similarity to the present probe. The correct match should appear in high ranks, ideally in rank 1.

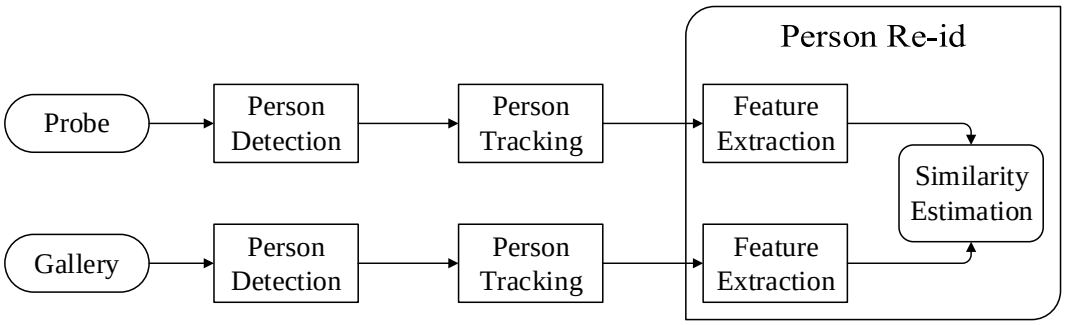


Figure 1.2: The general pipeline of an end-to-end person re-id system.

The general pipeline of a practical end-to-end person re-id system consists mainly of three modules: person detection, person tracking, and person retrieval [3]. It is currently widely agreed within the research community that person detection and tracking are separate vision problems [3, 5, 6]. Consequently, the re-id problem boils down into matching person bounding boxes across cameras. This in turn requires two main tasks to be performed, feature extraction and similarity estimation. Discriminative features must initially be extracted from person instances to describe each subject. A similarity measure should subsequently be employed to gauge the resemblance between the probe person and its gallery counterparts. This enables the gallery individual with the highest similarity to the probe to be deemed its correct match. In this manner, the probe person can be retrieved in another camera view, tracked until they leave that view, retrieved in a third view, and so on. A diagram depicting the general pipeline of a typical end-to-end person re-id system can be seen in Figure 1.2.

Relying on high-level biometric information that does not require specialist equipment such as face or gait recognition to solve the re-id problem is practically impossible for a number reasons [5]. Firstly, the data captured by surveillance cameras is usually of low resolution and poor quality which considerably complicates the face detection or gait extraction tasks. Secondly, many unconstrained factors such as viewpoint angle and pose variations yield notable changes in the way persons are perceived in different camera views. For instance, when a person is seen in a rear view, the back of the head is visible and not the face. Face recognition becomes unfeasible in this case. Finally, in real-world scenarios, self-occlusions between people or occlusions with other objects in the scene are very likely to happen. This has further implications on extracting reliable gait information. These reasons led researchers to heavily rely on holistic appearance models to tackle

the re-id task [7–12]; hence incorporating clothes, skin, and hair colours and textures.

1.2 Re-Identification Scenarios

Three different criteria can be considered when describing a person re-id protocol: (i) the nature of the data available, whether it is images or videos, (ii) the assumption made on the presence in the gallery set of a correct match to the query (probe) person, and (iii) the duration separating probe-gallery shots taken as this has considerable effect on the level of appearance change.

Considering the first criterion, four possible scenarios have been investigated in the literature so far:

- Each person is represented by one image in each camera view. In this case, there is a pair of images per person that need to be matched. This is known as single-shot person re-id.
- Each person is represented by multiple images in each camera view. These images could be taken at significantly separate time instances and thus do not constitute a video sequence. This is known as multi-shot person re-id.
- Video sequences (tracklets or tracks) are available for persons in both camera views. This is known as video-based person re-id. Some researchers refer equally to video-based re-id as multi-shot re-id [6, 13, 14]. To avoid confusion, when persons are represented by video sequences, i.e. frames taken at temporally consecutive instances, the video-based re-id terminology is used in this thesis.
- More recently, an additional re-id scenario is being considered where the probe person is represented by a single image (one camera view contains person images), and the gallery persons are represented by video sequences (the second camera view contains person tracklets). This is commonly known as image-to-video person re-id.

Among these four scenarios, the most extensively researched are the first three. Although a few methods have been recently proposed to tackle the image-to-video re-id problem [15–19], this research is still in its infancy. Its main application lies in cases where an image of a person of interest, e.g. a suspect, is available from a witness for instance, and this person needs to be traced in a specific camera network. This problem is significantly more challenging than other

re-id scenarios as it involves multi-modal data. Moreover, collecting task-specific benchmarking datasets is not very simple. In fact, the current datasets utilised are adapted from the video-based setting by randomly selecting a frame from each probe tracklet. The lack of suitable data added to the complexity of the task might have contributed into reducing the efforts made towards solving this problem so far.

Considering the second criterion, two possible cases are encountered:

- The person of interest shown in the query instance is also part of the gallery set. This implies that it is known in advance that at least one correct match exists for the probe when retrieval is attempted in a different camera view, and the algorithm's aim is to retrieve these occurrences. This is commonly known as closed-set or closed-world re-id.
- Alternatively, when it is not known *a priori* whether the probe person is part of the gallery set, the open-set or open-world re-id problem is faced. In this case, it is uncertain that at least one instance with the same identity as the probe exists in the gallery set. Subsequently, when a probe is recognised as not being part of the gallery, it is assigned a new ID for potential subsequent retrieval in other camera views.

In existing literature, the closed-world re-id scenario is far more researched than the open-world one [20–22]. In fact, when computing the similarity between probe and gallery elements, a threshold could be set to assert whether the most similar gallery person to the probe is similar *enough* to be deemed its correct match. Therefore, closed-world re-id should be solved first, and since the fundamental problem is yet to be solved, open-world re-id hasn't received a great deal of attention to date.

When it comes to the third criterion, also two cases are considered:

- The space and time separation between probe and gallery views are reasonably small. More precisely, the person instances to be matched are taken on the same day in relatively close locations. Although some changes in the apparel might still occur, such as zipped/unzipped jackets, carried objects and hair style, they will not be drastic. This scenario is known as short-term re-id.
- The space and time separation between probe and gallery views are significant. Person

instances could for example be taken on different days. In this case, the appearance change might be substantial such as a complete change in outfit. This scenario is known as long-term re-id.

Long-term person re-id involving a complete change in clothing is almost completely ignored in the literature so far as it is far more complex than short-term re-id. Although appearance might still change in the latter, since it is not substantial, an ideal short-term re-id system should be able to deal with such cases.

In the experimental work to follow in this thesis, the short-term closed-world re-id scenario is considered as it is currently the most active and relevant public datasets are available. As for the type of data used, the single-shot protocol is employed in Chapter 3 and the video-based one is explored in Chapters 4 and 5.

1.3 Challenges

Matching people across non-overlapping camera views underpins many challenges. While some are intrinsic to the re-id task, others are practical problems hindering real-world deployment. They could be summarised as follows.

- An intrinsic challenge to the task is cross-view appearance variations. By definition, a re-id system seeks to establish correspondences between person instances captured by two disjoint camera views. The concerned cameras often have different properties and parameters and are placed at various angles and elevations. Moreover, the scenes captured could be illuminated by different lighting sources or include varied degrees of clutter. These factors contribute considerably into increasing the intra-class variability (differences between cross-view instances of the same person) making it often bigger than the inter-class variability (differences between instances of different people) which makes the problem very challenging. The confusion is further accentuated by clothing similarities between different subjects, especially in winter where most people tend to wear dark clothes. Additional minor appearance changes exhibited by people transitioning between different scenes include zipping/unzipping jackets or using/eliminating accessories such as sunglasses, hats or scarves. For this purpose, simultaneously discriminative and robust features must be de-

signed or learned to mitigate the effect of these challenging attributes while maintaining the discernibility between persons. Additionally, advanced machine learning algorithms should be developed to learn the camera-to-camera transitions allowing to match corresponding instances successfully. This topic will be further elaborated in Chapter 3.

- For image-based re-id, only a pair or a small number of images are available per person. Meanwhile, existing datasets usually involve hundreds of different person identities. A large number of classes with a limited number of instances per class complicates the task of learning a good discriminative model that can effectively reflect the per class variability ensuring its separation from other classes in the data. This is known as the Small Sample Size (SSS) problem [23].

- The major hurdles preventing real-world deployment include the lack of enough data from concerned camera views for model training, coupled with the limited generalisation ability of most re-id methods into other camera views. These are added to the high labour cost required to label enough data, when available, in order to train effective re-id models.

- Since person detection and tracking accuracy have a major effect on the re-id performance [3, 24], integrating person detection, tracking and re-identification into a single end-to-end system that can successfully perform the task from raw input data is another problem that needs to be addressed to ensure re-id systems are applicable in real-life.

1.4 Motivation and Objectives

There is an ever-increasing demand on automated surveillance applications motivated by the ubiquitous availability of CCTV cameras in cities due to their ease of access, cheap cost, and crucial role in ensuring the security of vital infrastructure and the safety of the people involved. More precisely, solving the re-id problem can substantially improve forensic applications such as crime investigation, prevention of terrorist attacks, and intelligence gathering. Another important application is related to the welfare of vulnerable people. For instance, it facilitates the search for a missing person such as a child or an elderly who got lost in the city.

The wide deployment of surveillance cameras leads to a huge amount of data being collected on a daily basis. Manually searching this data by human operators is extremely costly and inaccurate.

This has contributed into making person re-id a very active research topic at present, and motivated us towards exploring this topic further.

Despite the remarkable development the field had witnessed in the past decade, the re-id problem is yet to be solved. The applicability of currently available re-id systems to real-world scenarios is hampered by a multitude of factors. Insufficient accuracy, lack of efficiency, and unrealistic requirements for model training are only a few. In fact, although a big step has been made lately towards improving the re-id accuracy [3], this was often accompanied by a high computational cost and the necessity of large-scale annotated datasets to be readily available for training. This problem is particularly encountered with supervised deep learning algorithms that currently produce state-of-the-art results on most re-id datasets [25–32]. Being the most accurate, these models suffer from major drawbacks. Firstly, specialist hardware such as Graphics Processing Units (GPUs) and high memory capacities are required to run these models. Secondly, the results produced are highly dependent on the availability of a large amount of annotated data containing matched instances of people from the concerned camera views. This in turn is either unrealistic or requires a great deal of human effort.

Given the above analysis, the main aim of this thesis is to move towards bridging the gap between conventional re-id and real-world requirements. The first step towards the goal would be to identify the re-id scenarios that are more likely to occur in real-life. The task being a surveillance application, it is very likely for persons to be represented by video tracks rather than single images, thus making video-based re-id the most realistic protocol. A solution to this problem is gradually attained throughout different phases of the research conducted for the purpose of this thesis. It could be broken down into the following objectives:

- A significant amount of research has been devoted into designing visual features describing pedestrians for the re-identification purpose. Despite the progress achieved, the performance attained is not yet sufficient [3]. Therefore, our initial objective is to evaluate existing feature representations against the attributes causing cross-view appearance variations. Besides inferring the best set of features, this work will reveal the main reasons causing the performance to drop.
- Knowing which challenges to tackle from the previous evaluation, the second step would

be to design a robust descriptor for video-based person re-id. Image-based descriptors have been often exploited to extract spatial features from separate frames in a video sequence, thus ignoring any temporal correlation between consecutive frames. Motivated by promising prior works in the field [33, 34], our second objective is to propose a robust spatio-temporal person descriptor. In order to fully understand the potential of such a descriptor, it should be evaluated in a supervised setting.

- The final step would be to develop a video-based person re-id system that meets real-world requirements. Such a system should be able to accurately retrieve pedestrians without the need of extensively annotated pairwise data for prior training. It should also have a good generalisation ability to different camera views and domains. In brief, the proposed system should be able to advance state-of-the-art results at a reasonable computational cost and data requirements.

1.5 Thesis Outline

This thesis contains six chapters. **Chapter 2** examines the existing literature on person re-id, the conception of the problem as part of a surveillance application and its evolution into an independent task. The mainstream methods developed for each re-id scenario are presented. The theoretical background of the algorithms employed and the methods that are directly related to this work are also highlighted. This includes image and video-based hand-crafted feature representations, distance metric learning, and multi-shot matching.

Chapter 3 details the methodology and the experiments undertaken to evaluate existing hand-crafted image-based feature extraction schemes as published in [35]. Public datasets are extensively annotated for the various challenges they represent, and the evaluation is conducted in a systematic manner to aid discovering the strengths and weaknesses of existing methods. Most importantly, it also uncovers the major problems hindering accurate re-id.

Chapter 4 introduces a novel supervised video-based person re-id method. In conjunction with distance metric learning, a robust spatio-temporal descriptor is proposed and carefully evaluated. Extensive analysis of the algorithm and comparison against state-of-the-art methods are conducted. The results obtained are also thoroughly discussed. The contents of this chapter are

published in [36].

Chapter 5 presents a novel unsupervised video-based person re-id method as published in [37]. Leveraging the spatio-temporal descriptor developed in Chapter 4, a set-to-set distance measure is proposed. Extensive experiments that analyse individual components of the algorithm and its performance against state-of-the-art methods are presented.

Chapter 6 concludes the thesis and proposes promising directions for future work. The contributions to knowledge made are also summarised.

1.6 Summary

This chapter introduced the person re-identification problem highlighting its difference from other biometric applications. A discussion of the inherent challenges accompanying the task and the motivations for choosing person re-id as the subject this thesis were subsequently presented. The aims and objectives for the research project were set, and the outline of the thesis was finally summarised.

Chapter 2

Literature Review

2.1 Introduction

Intelligent video surveillance has become essential with the huge amount of video data supplied by an ever-increasing number of surveillance cameras. A core component of an intelligent surveillance system is multi-camera tracking. When people are the concerned targets, multi-camera tracking encompasses three major tasks: person detection, mono-camera tracking, and inter-camera tracking. One way to solve the latter task is through person re-identification. When the cameras in a network have non-overlapping FOVs, disconnected persons' trajectories are reconnected through re-id. In addition to facilitating multi-camera tracking, a person re-id algorithm can further retrieve a subject of interest presented in one camera, in another target camera view regardless of the trajectory followed in between. This is vital for applications where re-building the whole trajectory is unfeasible or unnecessary, rather specific locations are suspected and need to be checked. For instance when the person is to be traced across various disconnected camera networks in different locations. In short, solving the person re-id problem provides a solution for inter-camera tracking and beyond.

Spotting a human from afar, analysing their characteristics, and resolving their identity is an intrinsic property of the human brain, validated by our powerful visual recognition abilities. This apparent trivial task entails enormous computational complexity, justified by the fact that half of the nonhuman primate neocortex is dedicated to visual processing [38, 39]. People detection, feature extraction, and classification are viewed as analogous tasks in person re-id. The complexity

of the visual cortex of the human brain suggests the challenges associated with all vision tasks, including person re-id [39]. Although a better understanding of the brain could inspire powerful learning algorithms as it did with artificial neural networks, computer vision will possibly provide solutions for re-id [25–32] without the need of a detailed explanation of the human visual perception to be brought to light [39].

Given the importance of the re-id task, a host of algorithms have been proposed especially in the past decade. These methods are broadly described in this chapter with more focus on the works related to this thesis. For the sake of clarity, a methodology-based taxonomy is used to organise various methods in this literature review. Additionally, a chronological narrative is presented for each of the categories identified, whenever possible.

2.2 Multi-Camera Tracking and Person Re-Id

Person re-id was previously referred to as inter-camera tracking and considered a module in a multi-camera tracking system [3]. The person re-id term was introduced in 2006 [40]. This is when the task started being recognised as an independent vision problem [3]. Despite the current independence of person re-id from multi-camera tracking, this section will examine the most prominent multi-camera tracking methods that are closely related to re-id, highlighting the reasons this separation was later evident.

The works that were more popular in early stages of the research suggest the use of inter-camera relationships based on spatio-temporal reasoning to facilitate multi-camera tracking. These methods could be generally divided into two main groups. The first group utilise camera topology and calibration techniques to match subjects' tracks across cameras by spatio-temporal analysis alone, and the second group incorporate appearance and visual cues to aid the matching, especially when cameras have non-overlapping FOVs.

Camera topology analysis aims at identifying the spatial and temporal relationships between different cameras in a network. This involves identifying any overlap or adjacency between camera views, and modelling the transition times from one camera to another [41]. As for multi-camera calibration, the goal is to map various cameras in a network into the same coordinate system. Once done, 2D coordinates of people in image planes are mapped into their 3D world coordinates in the

common coordinate system, thus enabling geometric analysis of their trajectories. These two tasks usually complement each other and are inextricably intertwined since resolving one yields a straightforward solution to the other.

Knowing the topology of a camera network could be beneficial in discarding unrealistic correspondences from the gallery set thus reducing its size which aids re-id [42–44], or can be used to establish correspondences between cross-camera tracks by spatio-temporal reasoning alone [45, 46]. Makris *et al.* [45] tackled the problem through learning a probabilistic model of the entry/exit zones of each camera and time transitions between cameras. A topological model could thus be inferred and used to calibrate the cameras in the network. Combined with temporal correlations, a tempo-topographical model is obtained and used to track individuals across the network. In [46], camera calibration parameters and camera topology are simultaneously learned through Maximum A Posteriori (MAP) estimation which allows the estimation of target individuals' trajectories across non-overlapping cameras. These methods rely solely on spatio-temporal reasoning to link tracklets of the same person in different camera-views and do not integrate any visual cues. The performance of such methods suffers when the velocity of different persons is not constant across the same path which is very likely in crowded scenes, or when the scene structure is complex and FOVs are disparate. On the other hand, knowing the 3D world coordinates of people through camera calibration allows the extraction of useful higher-level features to complement the visual descriptors built upon persons' appearances. These features may include the subject's height [47, 48], build [48], axes of symmetry [49], or body parts [47].

The other group of methods combine space-time cues with appearance models to establish tracklet correspondences. A leading model in this category was proposed by Javed *et al.* in 2008 [42], it integrates space-time variables such as velocity, entry/exit locations, and inter-camera transition times with a learned subspace of Brightness Transfer Functions (BTFs). A BTF [50] maps a colour observed in one camera view to its corresponding colour in another camera view. This mapping between two specific cameras is learned from labelled training data. In [44], the spatial and temporal topology of a camera network is learned by correlating activities in semantic regions obtained by decomposing each camera's FOV. Temporal and causal relationships are subsequently learned by canonical correlation analysis and combined with simple colour histograms to facilitate re-id. Despite some changes in appearance between various camera views, leveraging visual information

can improve inter-camera tracking accuracy.

In addition to the aforementioned limitations of these methods, calibrating all the cameras and estimating the topology of a large camera network is often time-consuming and unfeasible [41]. This has led to a bigger focus on the vision related aspects of re-id within the research community. Therefore, person re-id is treated as a pure computer vision problem in this thesis, and a solution is attempted without assuming any knowledge on the spatial position and parameters of the cameras concerned to ensure optimal generalisation ability.

2.3 Image-based Re-Id

Whilst video sequences are usually provided by surveillance cameras [33, 51], there are cases where only a few frames [24, 52], or possibly only one frame [7], could be retrieved from each view. This is expected in crowded, cluttered and noisy scenes where tracking is challenging. In that case, single or multiple images are used for re-identification, which is known as image-based person re-id [3, 53].

2.3.1 Feature Extraction

Upon the release of the challenging single-shot Viewpoint Invariant Pedestrian Recognition (VIPeR) dataset in 2008 [7], person re-id research exhibited near total separation from multi-camera tracking. Although re-id remains an important component of the latter, it extends to other applications where cameras are largely spatially separated and full trajectories of the people are not recoverable. Person detection and tracking being widely considered separate vision problems, the first building block of a typical re-id system is person description. In image-based re-id, this involves extracting features from person images describing their visual content. These features should allow distinguishing images of different people while highlighting shared features between images of the same person.

A brief description of the general feature extraction task can be given as follows. For simplicity, suppose we have a digital image of size $W \times H$ pixels stored by it's red, green and blue (RGB) colour values. A 3-dimensional matrix of size $H \times W \times 3$ is thus obtained. The values stored in this matrix undergo a certain mathematical analysis, depending on the problem at hand, to identify the

patterns they follow. These patterns could describe information related to texture, shape, colour, etc. The heads being mostly unidentifiable in low quality person images captured by surveillance cameras, relying on the appearance of the person dominated by their clothing or carried objects became imperative in person re-id. Therefore, most of the descriptors throughout the years rely on colour and texture information [7, 9–12, 54–56]. The most prominent are briefly described in the following.

Extracting holistic features from a person image is suboptimal for re-id. In fact, it is important to preserve some spatial information that enables comparing similar body parts together during the matching phase, whilst accounting for potential misalignment. For this reason the vast majority of hand-crafted person descriptors rely on part-based models. Although existing algorithms could be employed to automatically detect different body parts [57, 58], most of the re-id descriptors opt to approximate body parts by dividing the person image into a number of equally sized overlapping horizontal stripes [7, 11, 12, 59]. Since person images are usually vertically aligned [8, 12], a predefined part-model will significantly reduce the computational complexity.

Gray *et al.* [7] divide the person image into 6 overlapping horizontal stripes, and compute 16-bin histograms on 8 colour channels and the responses of 21 Gabor and Schmid texture filters. The features are extracted separately for each stripe and concatenated to obtain the final descriptor of the image. Following this work, the Symmetry-Driven Accumulation of Local Features (SDALF) algorithm was proposed in 2010 [8]. After human silhouettes are extracted, a vertical axis of symmetry and two horizontal axes of asymmetry are detected, thus partitioning the human silhouette into head, torso, and legs. The head is discarded, and weighted colour histograms are then extracted from the remaining body parts. More weight is assigned to the pixels closer to the vertical axis of symmetry aiming to mitigate background clutter and achieve some pose invariance. Subsequently, Maximally Stable Colour Regions (MSCR) are extracted and Recurrent High-Structured Patches (RHSP) are selected. For increased robustness against photometric changes , Yang *et al.* [60] developed the salient colour names descriptor by defining 16 colour names, and proposing a probabilistic model to assign pixel values into a specific colour name. A histogram of colours and colour names is then computed for each body-part (horizontal stripe) and concatenated. A more robust representation is obtained by fusing features extracted from 4 colour spaces. Another part-based model is proposed in [61]. It involves extracting joint colour histograms from 5 colour

channels in addition to Histogram of Oriented Gradients (HOG) [62] features. A kernel is also used to assign more weight to the central pixels in order to mitigate background effects.

As the use of colour information is evident in re-id, the main differences between various feature extraction schemes involve the choice of colour channels and texture descriptors. Moreover, the way local information is exploited may vary. For instance, an alternative to utilising horizontal stripes is computing features in small overlapping patches. Successive works by Zhao *et al.* [10, 59, 63, 64] compute dense colour and texture features in small overlapping patches. Specifically, for each patch, a 32-bin colour histogram is extracted from the LAB colour channels at 3 different scales. Moreover, a 128-dimensional Scale-Invariant Feature Transform (SIFT) [65] feature is computed in the same colour space. Patch-wise feature vectors are concatenated to form a high-dimensional image feature. Xiong *et al.* [54] employ the popular multi-scale uniform Local Binary Pattern (LBP) [66] texture feature in addition to 16-bin colour histograms computed on 8 colour channels as in [7]. An effort similar to [60] was also made by Zheng *et al.* [24]. The colour names descriptor introduced in [67] is computed locally in each of the small non-overlapping patches and average-pooled over the pixels in the patch. A Bag-of-Words (BoW) model is then employed to obtain the global vector by generating a codebook using the k-means algorithm. Ma *et al.* [9, 68] developed a descriptor based on Biologically-Inspired Features (BIF). This method is inspired by the human brain function where a hierarchical model is adopted [69]. It involves 24 multi-scale Gabor filters computed on the HSV colour channels separately, succeeded by max-pooling operations. Subsequently, the pixel's spatial location and neighbourhood information is summarised in a vector used to compute covariance matrices in small patches.

To simultaneously exploit both body-part and local information using horizontal stripes and patches, Liao *et al.* [11] extract joint colour histograms from the HSV colour channels in each patch, and combine them with Scale-Invariant Local Ternary Pattern (SILTP) [70] texture features. To achieve a multi-scale representation, the same features are also extracted from two down-sampled images by local average-pooling operations. Viewpoint invariance is achieved by max-pooling patch-wise features along the same horizontal stripe, thus maximising their local occurrence. A similar approach is proposed by Matsuwaka *et al.* [12]. Images are divided into 7 overlapping horizontal stripes and small patches. Patch-wise gradient and colour information is modelled using Gaussian distributions. Patches are weighted similarly to [8] and those belonging to the same horizontal

region are in turn summarised by another Gaussian distribution.

Other descriptors used in the literature are slight modifications of the aforementioned [55, 56, 71–74] and thus they will not be detailed here. We also note that the methods in [7, 10–12, 54, 68] will be explained in more detail in Section 3.2.1 as they are employed in our experiments.

In addition to low-level features, leveraging mid-level features such as semantic attributes has seen some attempts [75–77]. These methods require extensive attribute labelling for pedestrian images which is time-consuming and tedious. Layne *et al.* [75] introduced fifteen semantic binary attributes relevant to person re-id. Twelve of these are related to attire such as shorts, sandals, backpack, sunglasses; and three are soft biometrics including long-hair, short-hair, and gender. The authors manually annotate the attributes and train Support Vector Machines (SVMs) to detect them. A weighting is then learnt for each attribute highlighting its detection accuracy, discriminative power and relevant importance. This method is then fused with SDALF [8] for pedestrian matching. Other methods in the literature learn attributes from publicly available datasets such as internet images [76] or fashion photography datasets [77] and propose methods to transfer the knowledge into the re-id task.

The majority of these methods are mainly used to provide auxiliary information that aids conventional re-id by fusing attribute knowledge with low-level features. When tested on their own, their performance is not very competitive. Following the release of large-scale attribute-annotated datasets [78], subsequent methods employing semantic or data-driven attributes were mainly based on deep learning models [79–82].

2.3.2 Dimension Reduction

The hand-crafted features extracted in person re-id often involve thousands of dimensions. This is usually caused by the concatenation of patch-wise feature vectors in order to retain the spatial information. Following feature extraction, the similarity/dissimilarity between each query and its gallery counterparts is estimated. For this purpose, a distance metric is usually employed such as the Euclidean distance, or a learned distance that uses the latter at some stage (this concept will be explained in detail in Section 2.3.3). Subsequently, a nearest neighbour search principle is employed where the closest element to the query is expected to be the correct match.

Some research was once dedicated into studying the effect of high-dimensional data on nearest neighbour search and classification [83, 84]. The main outcome could be summarised as follows. Given a distance measure d, a query vector P and a gallery set $G = \{G_i | i = 1, ..., n\}$. We denote by $d_{\max}(P, G)$ and $d_{\min}(P, G)$ the maximum and the minimum distance obtained by comparing P to all the elements of G, respectively. Let m be the dimension of feature vectors P and $G_i, i = 1, ...n$. It has been proved in [83] that as m goes to infinity, a ratio known as the *relative contrast* goes to zero. That is,

$$\lim_{m \to \infty} \mathrm{E}\left[\frac{d_{\max}(P, G) - d_{\min}(P, G)}{d_{\min}(P, G)}\right] = 0, \tag{2.1}$$

where $\mathrm{E}[X]$ is the expectation of random variable X. In other words, in a high-dimensional space, the closest and the farthest points from the query become indistinguishable with respect to the minimum distance. That is, the points become almost uniformly distant from the query which makes the nearest neighbour notion lose its significance since the closest element found is not in fact the most "similar". This problem is referred to as the curse of dimensionality in the context of distance measures, and is more pronounced with the Euclidean distance.

Adding to the curse of dimensionality and important computational considerations, learning a distance metric in a high-dimensional space might be problematic. In fact, this task often involves matrix decomposition and/or inverting which might be costly or unfeasible when the dimension is significantly high. With a limited number of training data points and high-dimensional vectors, the matrices tend to become singular.

For this purpose, some works resorted into reducing the feature dimension before conducting the matching. To this end, the vast majority of methods employed Principal Component Analysis (PCA) [71, 85, 86]. Noting that the unsupervised re-id method developed in this thesis (Chapter 5) also employs PCA, we thoroughly describe it in the following.

Principal Component Analysis

Principal Component Analysis (PCA) [87] is a widely used unsupervised linear dimension reduction technique. It aims to transform a set of possibly correlated variables into a smaller set of uncorrelated variables. This is achieved by computing an orthonormal basis of the projection subspace where maximal variance is preserved. Formally, let $X = [x_1, ..., x_N]$ be the features matrix including N d-dimensional data points, the aim is to find projection directions, also called prin-

cipal components, which are the vectors forming matrix $W = [w_1, ..., w_k]$, where the projected features $W^T x_1, ..., W^T x_N$ retain maximal variance ($(\cdot)^T$ is the matrix transpose). Finding the first vector w_1 boils down into solving this problem:

$$\underset{||w_1||=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} [(x_i - \mu)^T w_1]^2, \tag{2.2}$$

where $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$ is the mean. By defining the covariance $\Sigma = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^T$, the problem in 2.2 can be written as

$$\begin{aligned} \underset{||w_1||=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} [(x_i - \mu)^T w_1]^2 &= \underset{||w_1||=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} [(x_i - \mu)^T w_1]^T (x_i - \mu)^T w_1 \\ &= \underset{||w_1||=1}{\operatorname{argmax}} \frac{1}{N} \sum_{i=1}^{N} w_1^T (x_i - \mu)(x_i - \mu)^T w_1 \\ &= \underset{||w_1||=1}{\operatorname{argmax}} w_1^T \Sigma w_1. \end{aligned} \tag{2.3}$$

w_1 can subsequently be found by eigen-decomposition of Σ. It is in fact the eigenvector corresponding to the largest eigenvalue of Σ. Similarly, $w_2, ..., w_k$ are the eignvectors corresponding to the remaining top eigenvalues of Σ. The orthogonal basis of the projection subspace W can thus be found and used to project d-dimensional data points into k dimensions where $k < n$.

2.3.3 Similarity Estimation

Despite the significant advancement in hand-crafted feature design [11, 12], no single set of features is able to successfully tackle all aspects of cross-view appearance changes while maintaining a strong discriminative power and the ability to generalise to different re-id datasets and scenarios. Therefore, while some researchers invested in designing robust features, others were devoted into developing matching methods that could contribute into mitigating re-id challenges and advancing the accuracy. The most popular among these methods are based on distance metric learning or subspace learning. We show in the following how distance metric learning and subspace learning are interconnected.

The most common type of distance metrics learned in re-id are Mahalanobis-like metrics [3]. Formally, given two column feature vectors x_i and x_j, the squared distance between them is given

by:

$$d_M^2(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j), \tag{2.4}$$

where M is a Positive Semi-Definite (PSD) matrix ($M \succeq 0$) to be learned. Being positive semi-definite, M can be decomposed into $M = W^T W$. Subsequently,

$$d_M^2(x_i, x_j) = (x_i - x_j)^T W^T W (x_i - x_j) = [W(x_i - x_j)]^T W(x_i - x_j) = ||Wx_i - Wx_j||^2, \tag{2.5}$$

Therefore, computing the Euclidean distance between vectors Wx_i and Wx_j, which are the projections of x_i and x_j into the subspace W, is equivalent to computing the distance d_M between vectors x_i and x_j using the learned matrix M. As a result, Mahalanobis-like distance metrics and subspace learning methods that employ the Euclidean distance are in essence similar. They both aim to learn a subspace where positive examples (images of the same person) are pulled closer to each other while negative examples (images of different people) are pushed apart. This is illustrated in Figure 2.1. While some methods explicitly learn the distance matrix M (Equation 2.4), others learn the subspace projection matrix W (Equation 2.5) depending on the nature and tractability of the optimisation problem formulated. It is also worth noting that learning the distance metric requires pairs of labelled data points, vectors x_i and x_j associated with a label y_{ij} indicating whether they belong to the same person or not. In person re-id, when these points represent the same person from different camera views they are known as equivalence constraints [88]. Alternatively, when they represent different people they are known as inequivalence constraints. Together they form the pairwise constraints [88].

Distance metric learning was conceived long before re-id [89] and used in various applications such as clustering [89–92], classification [91–94], retrieval [95–97], face verification [98, 99] and many more [100]. We highlight in this section the leading methods that were mainly designed and used for the re-id purpose.

One of the early distance metric learning methods in re-id is Probabilistic Relative Distance Comparison (PRDC) [101]. Zheng *et al.* propose in this work to learn a distance that maximises the probability of a pair of positive examples having a smaller distance than a pair of related negative examples. This is motivated by the idea that intra-class and inter-class variations can be different

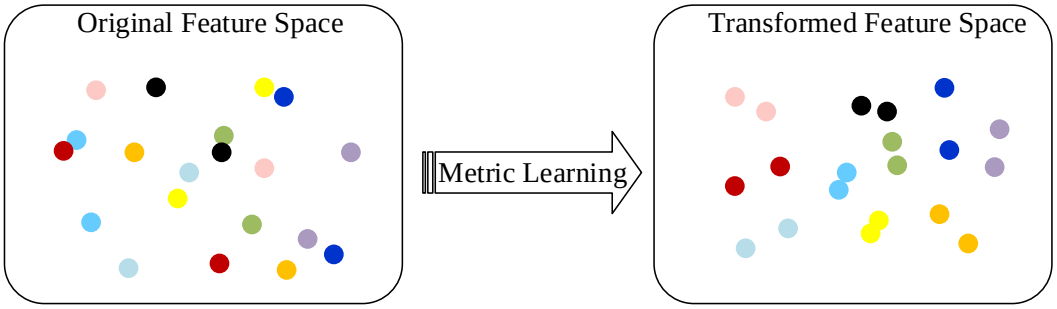


Figure 2.1: An illustration of distance metric learning. The goal is to learn a subspace where positive examples are closer to each other and negative examples are farther apart. Person instances holding the same ID (positive examples) are represented using the same colour.

between pairs of images for various reasons (e.g. occlusions, viewpoint, pose, illumination), therefore using a probabilistic model will tackle these variations while preventing over-fitting when only scarce annotated data is available.

Following this work, a few other distance metrics were developed. Hirzer *et al.* [102] propose to relax the constraint of M being a PSD matrix which yields a very simple and efficient solution to the problem that also produces good results. Mignon *et al.* [55] propose the Pairwise Constrained Component Analysis (PCCA) algorithm that learns a linear function used to map high-dimensional data into a low-dimensional subspace using pairwise constraints reflected by an objective function. The kernel trick is then incorporated into this method to learn the non-linearities, yielding kPCCA.

An early distance metric that witnessed great success in re-id is Keep It Simple and Straightforward Metric (KISSME) [85]. Using the space of intrapersonal differences formed of vectors $z_{ij} = x_i - x_j$ where x_i and x_j belong to the same person, and extrapersonal differences ($z_{ij} = x_i - x_j$ where x_i and x_j belong to different people), the authors propose to leverage the log-likelihood ratio test with the null-hypothesis H_0 that a pair is dissimilar and its alternative H_1 that a pair is similar. It is given by:

$$d(z_{ij}) = 2\log\left(\frac{p(z_{ij}|H_0)}{p(z_{ij}|H_1)}\right). \tag{2.6}$$

These two spaces are assumed to follow zero-mean Gaussian distributions. Consequently, the

probabilities in Equation 2.6 can be calculated as:

$$p(z_{ij}|H_0) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_E|}} \exp\left(-\frac{1}{2} z_{ij}^T \Sigma_E^{-1} z_{ij}\right), \tag{2.7}$$

$$p(z_{ij}|H_1) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_I|}} \exp\left(-\frac{1}{2} z_{ij}^T \Sigma_I^{-1} z_{ij}\right), \tag{2.8}$$

where $|\cdot|$ is the matrix determinant operator, d is the dimension of z_{ij}, and Σ_I and Σ_E are the respective covariance matrices of the intrapersonal and extrapersonal difference spaces. By computing the $\log$ and eliminating the constants, this equation simplifies into:

$$d(z_{ij}) = z_{ij}^T(\Sigma_I^{-1} - \Sigma_E^{-1})z_{ij}. \tag{2.9}$$

Therefore, the matrix M is found to be equal to $\Sigma_I^{-1} - \Sigma_E^{-1}$ that can be directly computed from the training data. M is then projected into the cone of PSD matrices by eigenalaysis, i.e. by replacing negative eigenvalues with a small positive number and reconstructing M. Moreover, PCA is applied in [85] to reduce the features' dimension before learning the metric. The popularity of KISSME is gained from its simplicity through preventing complex iterative optimisation algorithms and its promising accuracy. On the other hand, a drawback of KISSME is the need to reduce the dimension of the feature vector before learning to avoid matrices singularities, which might cause the loss of some useful information.

Subsequent methods mainly optimise a Fisher-type criterion to learn the projection subspace [11, 54, 71]. They aim to maximise the between-scatter S_b and minimise the within-scatter S_w. For the original Fisher Discriminant Analysis (FDA) [103], these matrices can be computed as follows. Let μ be the mean of all data points, μ_c the mean of points in class c, N_c the number of elements in class c, and C the total number of classes. if x_c^i is the i^{th} data point in class c, S_b and S_w are given by [71]:

$$S_\mathrm{b} = \sum_{c=1}^{C} n_c(\mu_c - \mu)(\mu_c - \mu)^T, \tag{2.10}$$

$$S_\mathrm{w} = \sum_{c=1}^{C} \sum_{i=1}^{N_c} (x_c^i - \mu_c)(x_c^i - \mu_c)^T. \tag{2.11}$$

The Fisher criterion is then derived as:

$$\mathcal{J}(w) = \frac{w^T S_{\mathrm{b}} w}{w^T S_{\mathrm{w}} w}, \tag{2.12}$$

where w is a column of the subspace projection matrix W to be learned, that maximises the Fisher criterion. This problem can be cast into the following:

$$\max_{w} w^T S_{\mathrm{b}} w, \quad s.t. \quad w^T S_{\mathrm{w}} w = 1. \tag{2.13}$$

Assuming the data in each class follows a Gaussian distribution, the solution to this problem can be derived by eigenvalue decomposition of $S_{\mathrm{w}}^{-1} S_{\mathrm{b}}$. An optimal projection direction w is simply an eigenvector associated with a large eigenvalue. These column vectors are used to form the projection matrix W.

To preserve the localities in the data, Local Fisher Discriminant Analysis (LFDA) was then proposed in [104] and adapted to person re-id in [71]. Different from FDA, LFDA learns local embeddings in a manner similar to Locality Preserving Projections (LPP) [105] by multiplying the scatter matrices by an affinity matrix so that closer points in the same class are given more weight than points that are farther apart. The problem is subsequently solved similarly to FDA.

In a similar spirit, Marginal Fisher Analysis (MFA) was proposed in [106]. The main changes it makes on FDA are also the scatter matrices. It is a graph embedding method that employs within-class and between-class nearest neighbours in the scatter information, in the aim of relaxing the assumption on the distribution of the data and increasing the separability between classes. The problem obtained can also be solved by general eigenvalue decomposition.

Kernelised versions of LFDA and MFA were used for the re-id purpose in [54], outperforming their non-kernelised counterparts. Also an improvement of PCCA was proposed in this work by adding a regularisation to avoid over-fitting, and denoted rPCCA. The kernel trick ensures non-linearities can also be learned by implicitly mapping the data into a higher dimension space where it exhibits better separability.

A problem these methods might encounter is the singularity of S_{w} when the feature dimension is high and/or the sample size is small. This is usually solved by carefully adding a regularisation

parameter to the matrix diagonal entries. To account for this problem, the Null Foley-Sammon Transform (NFST) was proposed in [107] and employed in re-id in [108]. Instead of maximising the Fisher criterion, NFST aims at learning a discriminative subspace where the within-scatter is zero and the between-scatter is positive. This ensures data points belonging to the same class are collapsed into a single point in the learned subspace, while different classes remain separable. This is achieved by learning projection directions w that satisfy the following conditions:

$$w^T S_{\mathrm{w}} w = 0 \quad \text{and} \quad w^T S_{\mathrm{b}} w > 0. \tag{2.14}$$

This problem also has a closed-form solution that can be efficiently computed by eigenvalue decomposition. In fact, the projection directions sought are found to lie in the space shared between the null-space of S_{w} and the orthogonal complement of the null-space of S_{t}, where S_{t} is the total scatter matrix given by $S_{\mathrm{t}} = S_{\mathrm{b}} + S_{\mathrm{w}}$. A kernel can also be integrated into NFST yielding kNFST. The superiority of kNFST compared to previous methods is two-fold. Firstly, it ensures maximal separability of the data by collapsing each class into a single point. Secondly, It has no free parameters to tune since no regularisation is needed to avoid matrices singularities, which makes it more suited for small size data.

Another very popular distance metric in person re-id is Cross-View Quadratic Discriminant Analysis (XQDA) [11]. XQDA brings a slight modification to KISSME that alleviates the need of reducing the feature dimension before learning the distance matrix. In fact, Liao *et al.* [11] argue that the unsupervised dimension reduction step (PCA) employed by KISSME before learning the metric is not optimal. Therefore, they suggest learning a subspace projection matrix W in a supervised manner prior to learning the metric M. W is efficiently found by optimising a Fisher-type criterion similar to the one in Equation 2.12, and M is computed similarly to KISSME. Consequently, the features are projected into the low-dimensional space before being compared using M. More details on XQDA can be seen in Section 3.2.3.

Before the predominance of Mahalanobis-like distance metric learning, various learning algorithms were attempted in re-id. Gray *et al.* employed the AdaBoost algorithm [7]. It iteratively searches the feature space for the best features and combines the classifiers found to obtain the final similarity function. Variants of SVMs were also utilised for the task. An Ensemble of RankSVMs

was learned in [109] on small overlapping groups of the training data and combined using boosting to form a strong ranker. In [110], a sample-specific SVM was learned for each individual followed by a dictionary pair and a mapping. Each probe image is then mapped into a weight vector based on its visual features. The main drawback of these methods is their lack of efficiency compared to distance metric learning.

2.3.4 Deep Learning Methods

Deep learning systems achieved remarkable success in many computer vision tasks including object detection, image classification, face recognition, and many more [111]. However, it was not until very recently that similar success has been achieved in person re-id [25–32]. Knowing that such systems can learn features and metrics simultaneously in an end-to-end fashion, the availability of enough labelled training data is a necessity. When the latter started becoming available, a variety of deep learning systems have been introduced [52, 112–117]. These can be generally divided into two major types: the Siamese networks and the classification networks. Siamese networks take image pairs, triplets or even quadruplets as input to learn the model. They include at least two identical sub-networks. A very basic Siamese network is shown in Figure 2.2. Meanwhile, the networks that approach the task from a classification perspective consider each person identity as a separate class that the network should learn.

The feature extraction sub-net in these networks consists mainly of one type or a combination of Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTMs). On the other hand, the matching layer could be for instance a pairwise, triplet or quadruplet loss or simply an off-the-shelf distance measure such as the Euclidean or Cosine distance.

The first few person re-id datasets released were small in size, and thus did not provide the means to learn deep models. In 2014, Li *et al.* [52] collected the CUHK03 dataset that includes 13,164 images of 1,360 pedestrians. Containing over a thousand different identities, CUHK03 was considered a large-scale dataset [3]. They also proposed one of the first deep learning re-id models. Briefly, it uses a Siamese type network that takes 2 input images, similar or dissimilar. A convolutional layer is first applied to each image succeeded by a patch matching layer that computes similarity scores between patches in the same horizontal stripe. Another convolutional layer with

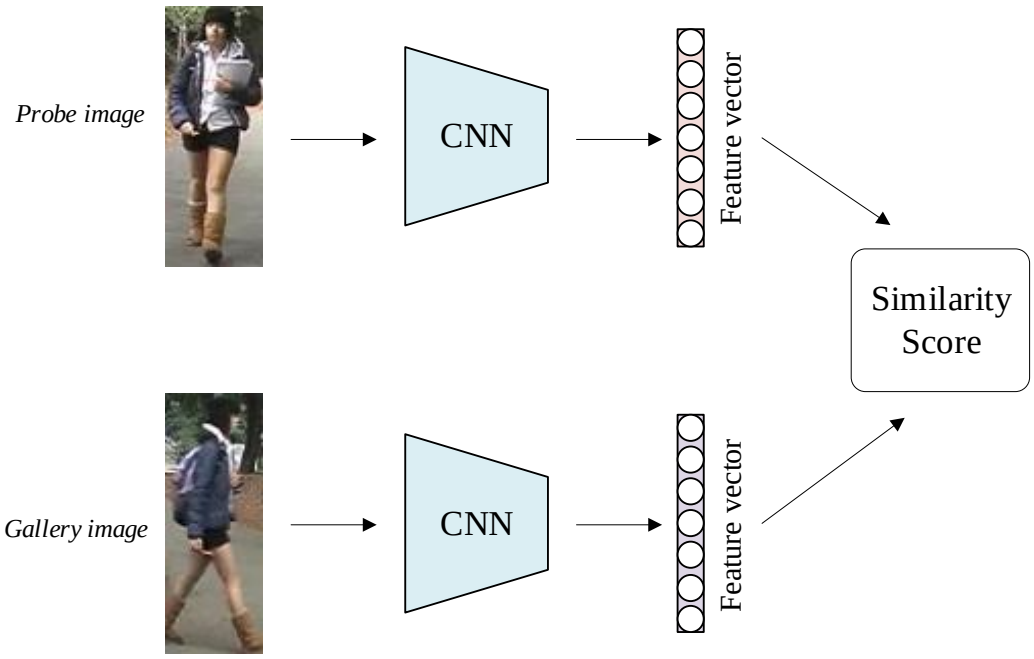


Figure 2.2: A simple pairwise Siamese network.

shared weights follows before a fully connected layer and a Softmax function. As more large-scale datasets were later released [24, 118], a plethora of deep learning algorithms were proposed. The most prominent published works are briefly summarised in this section.

A few ensuing works also employ the pairwise Siamese model. Ahmed *et al.* [112] improve the CNN model in [52] by computing bi-directional neighbourhood differences between corresponding feature maps of two input images in order to achieve some positional robustness. Varior *et al.* [113] integrate gating functions between CNN layers to emphasise local mid-level patterns, and Zheng *et al.* [119] embed modules that leverage spatial attention and enforce attention consistency between images of the same person.

Instead of considering similar and dissimilar image pairs separately, a straightforward alternative is to use triplets. A triplet consists of a query image, a gallery image with the same identity, and a another one with a different identity. Cheng *et al.* [114] propose a CNN part-based model that fuses holistic and local body-part features while improving on the conventional triplet loss function [120] by forcing the intra-class distance below a fixed margin. Subsequent methods using the triplet loss or its variants mainly modify the network's architecture to focus on multi-

scale features [121], attributes [122], attention [123], or improved similarity learning [124, 125]. Inspired by the triplet model, Chen *et al.* [126] propose a quadruplet loss by adding an extra negative pair to each triplet with the intent of ensuring the intra-class distance is always smaller than the minimal inter-class distance regardless of the probe image in question.

In our collaborative work published in [127], the triplet loss was further improved by replacing the standard Euclidean distance with a weighted Euclidean distance in the loss formulation. The weights are dynamically computed based on the variance of each feature in a given batch. Features exhibiting higher variation are assigned more weight as they are deemed more discriminative. Also the feature extraction sub-net is improved by integrating channel attention using Squeeze and Excitation (SE) modules [128] into the ResNet-50 [129] backbone to emphasise more important features.

These models are particularly useful when available data is limited per identity as only a pair of images is required per subject. With larger datasets, the classification model is more promising. It leverages multi-label information of different classes rather than relative positive and negative pairwise labels regardless of their class.

A large amount of work was done along this line in the past couple of years. The major innovations made contribute into improving the robustness by using deformable part-models in contrast to earlier works employing rigid part-models (fixed horizontal stripes) [114, 130]. For instance, Zhao *et al.* [115] localise different body regions using existing body joints detection methods of which features are extracted and fused with full body features. A similar fusion approach is adopted in [131] but within a multi-scale network. Sun *et al.* [117] start with a uniform partition of the body which is then refined by enforcing consistency within each part. In a later work [29], the visibility of each region is learned and region-level comparison is performed between images to solve partial re-id where bounding boxes might cover only a part of the human body. In a similar spirit, an attention mechanism could be explicitly embedded into the network for the system to learn where to focus. Li *et al.* [116] propose the harmonious attention that combines regional, spatial, and channel attention. The network is thus able to locate discriminative pixels, regions, and feature channels. Part-based models achieved remarkable performance among deep-learning methods as they are able to effectively deal with pedestrian misalignment, i.e. when the person is not centred in the image.

Despite recent success, the limitations of deep learning systems reside in their high computational cost, added to the critical availability of enough labelled training data from the camera views involved. Despite the use of generative models and transfer learning techniques to augment and diversify the training data, or to transfer the knowledge from one domain to another [27, 132–136] the performance attained by these algorithms is still far below their fully supervised counterparts that produce state-of-the art results [25–32].

2.4 Video-based Re-Id

The recent popularity of video-based person re-id [31, 137–144] is motivated by two main reasons. Firstly, person videos are generally available from surveillance cameras which makes the video re-id problem a more realistic one. Secondly, video sequences include rich visual information and enable the use of temporal cues to design robust person representations that could circumvent cross-view appearance changes. This has led to a number of video-based person re-id methods being developed in the past few years [33, 34, 139–145]. We describe the leading algorithms with more focus on those related to our work in this section.

2.4.1 Person Description

Since a video sequence is merely a number of successive frames, extracting image-based features from separate frames and aggregating them is an intuitive way to represent a video tracklet. This approach has proved in fact popular for the video re-id problem especially with the advancement achieved in image-based feature design [53, 141, 146–148]. However, frames in a video sequence are not a set of independent images, they are rather temporally correlated. Exploiting this temporal relationship can potentially describe some sort of motion information that would complement the visual cues to aid re-id. For this purpose, various ways of integrating temporal information into video-based person description were explored.

Early methods build appearance models from separate multiple shots without explicitly leveraging any temporal information. They either assign a single signature to each individual or compute multiple representations by grouping similar frames together through clustering or other procedures [14, 33, 34, 137]. In this case, more effort is dedicated into designing appropriate multi-shot matching methods. In this section, we cover works that propose new person descriptors for the

video-based scenario, rather than those using classical image-based features to design matching methods. The latter will be examined in a subsequent section (Section 2.4.3).

In the very first work on video re-id, Gheissari *et al.* [40] develop a spatio-temporal segmentation algorithm to detect stable foreground regions and produce a salient edgel histogram in each region. Additional visual cues including normalised colour histograms from two colour channels are added to the edgel histograms to form the person's signature. A triangulated graph that localises body-parts is then fit to the image for matching.

In a more recent work, Wang *et al.* [33, 149] incorporate temporal information directly into the person representation. Fragments representing walking cycles are initially extracted from person sequences using the Flow Energy Profile (FEP) signal. Each fragment is then divided spatially and temporally into cells according to human body topology and walking cycle phases. Histograms of Oriented 3D Gradients (HOG3D) features [150] are computed in each cell and concatenated. Mean colour values from six colour channels are also extracted from small overlapping patches and temporally averaged over each fragment. The two types of features are finally concatenated to form the fragment representation coined ColHOG3D.

Similarly, Liu *et al.* [34] also extract walking cycles after performing a correction on the FEP signal to reduce the noise. The fragments obtained are aligned and divided spatially and temporally into body-action units on which Fisher vectors describing colour, texture and gradient information are extracted and concatenated to obtain the final descriptor denoted STFV3D.

Dense temporal gradients are computed on small patches at six different scales in [151]. An energy map is then built for each sequence by eliminating low gradient values representing short-term motion and background noise through adaptive thresholding operations. The energy maps obtained are finally divided into patches and horizontal stripes before being vectorised and concatenated.

Khan *et al.* [145] construct a signature for each person sequence and call it Part-Appearance Mixture (PAM). After low-level features are extracted from separate frames, Gaussian Mixture Models (GMMs) are used to model the features' distribution in the upper, lower and full body. Since GMMs do not fall in the Euclidean space, a distance measure is adapted from KISSME for the matching task.

Despite showing promising accuracy, the main drawbacks of existing spatio-temporal descriptors lie in their lack of efficiency and insufficient robustness to cross-view appearance changes. A successful re-id algorithm should be able to strike the right trade-off between invariance to cross-camera changes, discriminability, and computational efficiency.

2.4.2 Clustering

A common method to process frame-wise features for multi-shot ranking or set-based matching is by clustering them. The most widely used algorithm for this purpose is the k-means algorithm [152]. Since k-means is also employed in our work (Chapter 5), we detail it in the following.

K-means Algorithm

Given n data points $x_1, ..., x_n$ where $x_i \in \mathbb{R}^d$, k-means aims to partition this data into k disjoint clusters $C = \{C_1, ..., C_k\}$ by optimising a criterion J that minimises the within-cluster variance:

$$J = \underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2, \tag{2.15}$$

where $|| \cdot ||$ is the Euclidean distance and μ_i is the mean of points in cluster C_i. i.e. the cluster centroid.

After random initialisation of the centroids, the algorithm alternates between an assignment step and an update step until a stopping criterion is reached, usually when the clusters become stable.

- Assignment step: Each data point x is assigned to the cluster with the closest centroid, i.e.

$$\underset{C_i \in C}{\operatorname{argmin}} ||x - \mu_i||^2, \tag{2.16}$$

- Update step: The means of the new clusters are computed to obtain the updated centroids:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_i \in C_i} x_i, \tag{2.17}$$

  where $| \cdot |$ is the set cardinality operator.

When the assignments no longer change, each point will be associated with a label indicating the cluster it belongs to.

Initialising the centres of the k-means algorithm is critical for speeding up the convergence of the algorithm and obtaining accurate clusters. For this purpose, Arthur *et al.* [153] proposed the k-means++ seeding method for initialisation. It could be summarised as follows:

1. The first centre is picked uniformly at random from the data points.
2. The distance $d(x)$ of each data point x from the closest centre already chosen is computed.
3. The new centre is randomly selected from the remaining data points with a probability proportional to $d^2(x)$.
4. Steps 2 and 3 are repeated until the desired number of centres have been chosen.
5. The k-means algorithm is subsequently applied.

This initialisation method is used with the k-means algorithm employed in Chapter 5.

2.4.3 Matching Methods

Although designing problem-specific video-based person descriptors demonstrated superiority over off-the-shelf image-based features used in video-based re-id, the performance attained was less than satisfactory. Overcoming drastic appearance changes across cameras can better be achieved through learning. Therefore, another potentially promising way to exploit the rich information provided in a person tracklet is to design matching methods accordingly. These should be able to select or successfully fuse information from multiple instances (frames or video fragments) in order to attain an optimal representation. The methods developed for this purpose can be roughly divided into two categories: distance metric learning and multi-shot matching.

Most of the distance metrics designed for image-based re-id can be equally applied to the video-based problem by sampling for instance to obtain positive and negative pairs for model training. Nevertheless, a few algorithms were specifically devised for the video re-id task. You *et al.* [143] argue that the inter-class variation is smaller in videos compared to images since different people might have similar motion, therefore a stringent constraint is posed in their proposed distance metric to ensure maximal inter-class separation in the latent feature space. Zhu *et al.* [154] con-

sider two types of variation in their distance metric design, intra-video variation representing the variability among frames in a person's video, and inter-video variation describing the variation between different videos. Taking both types into account, a distance metric is formulated to learn two projection matrices simultaneously, one to reduce intra-video variation making each sequence more compact which enhances between-video separation learned by the second matrix.

Another type of methods aim to select and rank the best frames or fragments for matching. These are known as multi-shot matching methods. A ranking function was proposed with ColHOG3D descriptor [33]. The latter produces multiple representations for a person's sequence using video fragments (walking cycles) which in turn requires a set-based method for matching. To this end, the authors devise a ranking function that simultaneously learns the most discriminative fragments and the parameters of the ranking function.

Karanam *et al.* [13] formulate multi-shot re-id as a block sparse recovery problem where each probe image is a linear combination of corresponding gallery images in the embedding space learned by LFDA. Given this assumption, a dictionary including all gallery instances can be constructed, and each probe image will be a sparse linear combination of the elements of the dictionary. The block concentration of the coefficients determines the identity of the probe. In their later work [14], Karanam *et al.* cluster frame-wise features using k-means algorithm, and formulate the distance between vector sets as the minimum distance between their respective affine hulls. Ways to integrate existing distance metric learning methods into this framework were also discussed.

Alternative ways to clustering or randomly selecting frame-wise features were also developed. Cho *et al.* [137] estimate the pose of the pedestrians in each frame and divide them into four groups representing various pose angles covering 90° each. The matching is subsequently undertaken by learning to weigh pose pairs using an SVM. Huang *et al.* [138] segment the video sequence into fragments using occlusion information, and devise a method to measure the noise level of each segment using a self-paced outlier detection method. The set-based distance measure then employed weighs corresponding segments using their noise levels.

Other works that fall within this category include simple set-to-set distance measures that compute a dissimilarity score between two sets of vectors without any learning. The most popular among

these are the Minimum Point-wise Distance (MPD) also known as infimum distance [34], and the Average Point-wise Distance (APD) [155]. In both cases, all possible combinations of pairs of vectors from two sets are considered and their relative distances computed. The minimum value is selected for MPD and the average value is found for APD.

These algorithms provide efficient and simple ways to boost the performance for the multi-shot scenario in conjunction with robust person descriptors. However, various unconstrained factors such as occlusions and background clutter might affect fragment segmentation or pose estimation. Moreover, random frame selection might cause the loss of important information. For this purpose, we find clustering to be the optimal way to group similar frames based on multiple hidden modalities in the data.

2.4.4 Deep Learning Methods

The first work using deep learning in video-based re-id is published in [141]. McLaughlin *et al.* propose a simple pairwise Siamese network architecture. For each frame, three colour channels and two optical flow channels are fed into a CNN layer followed by a recurrent layer. The RNN layer can capture the information flow between time-steps. An average-pooling layer follows thus aggregating the features extracted from all the sequence frames into one global vector. A joint identification/verification loss is then optimised using both the cross-entropy and the Siamese loss functions. Xu *et al.* [142] improve on this model by inserting an attentive spatial pooling layer between the CNN and the RNN to capture meaningful regions in the image. The average-pooling layer is also replaced by an attentive temporal pooling layer. The latter uses a specific mechanism to weigh the frames in a sequence before aggregating them using a weighted average-pooling framework.

Attention models proved popular in deep learning methods for video-based re-id as they provide the means to identify important spatial regions in an image and informative frames within a sequence [139, 140, 142]. Similarly to [142], Liu *et al.* [140] also employ the idea of temporal attention by frame weighting. However, this is achieved by integrating a quality generation unit in the CNN model that generates a score for each frame based on its quality. Frame-level features are then combined using their quality scores. Li *et al.* [139] sample a few frames from each video sequence to avoid redundancy. They then combine multiple spatial attention models to

locate salient regions. Diversity is imposed on these models to ensure each one of them focuses on a different body part. The spatial features obtained are then fused using additional temporal attention methods.

Suh *et al.* [156] on the other hand tackle the misalignment problem by developing a two-stream network that separately generates part and appearance maps for each image. These are later aggregated by bilinear-pooling, i.e. by computing the local outer product at each location followed by spatial pooling. This framework ensures body parts representations are aligned which enables the use of a part-level similarity measure. Frame-level representations are simply average-pooled over a video sequence. In addition to attention and part-based methods, video re-id also witnessed the use of reinformcement learning. Zhang *et al.* [144] employ a sequential decision making process by successively feeding an agent pairs of probe and gallery images to which a reward or penalty is earned based on the decision made, i.e. whether the images are different, same, or unsure. Like so, an aggregation function and a similarity function are concurrently learned. The features used are extracted using traditional CNNs such as AlexNet by optimising a combination of three loss functions.

Semantic attributes were also recently exploited. Zhao *et al.* [26] disentangle the CNN frame-wise features computed into groups representing different semantic attributes such as gender, age, glasses, hat, etc. These are learned by binary classification and corresponding features are weighed based on the attribute's recognition confidence for temporal aggregation. On the other hand, Hou *et al.* [30] tackle specifically the occlusions problem by locating the occluded regions in a frame and recovering them using an autoencoder-type network. This is further informed by adjacent clean frames found by temporal attention.

In contrast to end-to-end methods, Zheng *et al.* [148] test their large-scale video dataset Motion Analysis and Re-identification Set (MARS) using a CNN feature extraction scheme known as ID-discriminative Embedding (IDE) [157] which is based on AlexNet [158]. The matching is performed separately using XQDA distance metric.

Similarly to hand-crafted systems, a better alternative to computing one global representation for the whole person sequence is breaking it down into smaller segments and employing a set-based method in the matching phase. Along these lines, Chen *et al.* [31] densely divide the video se-

quence into equal length segments called snippets taken at a time step of 4 frames. CNNs, LSTMs, and fully connected layers make up the architecture of the network that embeds the snippets after weighing their frames using a temporal attention mechanism. Only reliable snippets are considered in the final matching score. In Si *et al.* [159], frame-level features are extracted using a CNN. A dual-attention mechanism is subsequently employed by combining both intra-sequence and inter-sequence attention mechanisms for feature sequences refinement and alignment, respectively. Dense pairwise distances of refined features are aggregated by average-pooling.

Just like their image-based counterparts, these methods require a large amount of annotated data and suffer from substantial computational cost. When tested on small datasets, they are often pre-trained on large-scale benchmarks such as MARS and fine-tuned on the target dataset to achieve competitive performance [148].

2.5 Image-to-Video Person Re-Id

Image-to-video person re-id encompasses interesting real-life applications related to forensics and the well-being of vulnerable people as discussed in Section 1.2. Although the problem hasn't received similar attention compared to other scenarios, a few attempts were recently made [15–19].

Zhu *et al.* [15] tackle the problem from a dictionary learning perspective. A feature projection matrix and a pair of heterogeneous dictionaries are jointly learned. Video features are projected into a subspace where they exhibit higher compactness, and the final representations of images and videos are compared using their coding coefficients in their respective dictionaries. In [16], frames are segmented into regions based on their colour information and a salience value is computed for each region. The most salient regions are selected and clustered before using a point-to-set distance measure for matching.

The few remaining methods rely on deep learning to solve this task. Zhang *et al.* [17] design an end-to-end network that simultaneously learns the feature representation and the similarity measure. The feature extraction sub-net involves CNNs for images and CNNs followed by LSTMs for videos to embed space-time information. Weighted average-pooling is finally applied to the LSTM output to emphasise more useful frames before being passed to the similarity sub-network

for matching. A similar type of architecture is employed in [18] where a body-parts attention unit is additionally incorporated into the network with a different loss function. Similarly, Wang *et al.* [19] also propose an end-to-end network. However, different from [17], only CNNs are used for feature extraction and a domain alignment sub-network is added before the similarity sub-network. The latter is a k-NN triplet loss rather than a conventional one. The difference between the two is that the first additionally uses nearest neighbour information to mine the least noisy video frames to be input into the loss function.

With the remarkable progress towards solving conventional re-id lately, the image-to-video protocol is attracting more and more attention [15–19]. Nevertheless, a lot of room still exists for improvement in terms of accuracy and real-life applicability. The main bottleneck towards achieving this goal is the lack of true cross-modal data.

2.6 Evaluation Metrics

The evaluation metrics employed to measure the performance of person re-id systems are presented in this section. As these might differ with the re-id scenario involved, whether it's open-world or closed-world for instance, only the metrics associated with the closed-world problem used throughout this thesis will be explained in the following.

2.6.1 Cumulative Matching Characteristic

Person re-id is generally viewed as a ranking problem. When a probe instance is presented to the system, a ranking of the gallery elements is generated based on their similarity to the given probe. The item with the highest similarity appears in rank 1. In a typical closed-world scenario, at least one correct match (ground truth) exists in the gallery set to each probe. Whenever a probe instance is given and a ranked list is generated, the rank of its correct match is recorded. If there are more than one, the highest rank of a correct match is considered.

The Cumulative Matching Characteristic (CMC) at a given rank k is then computed as the percentage of probe instances that had a ground truth appear at a rank higher than or equal to k [5]. For instance, CMC(@5) is the percentage of queries that had a correct match appear at a rank higher than or equal to 5, i.e. from 1 to 5. Although CMC(@1), also known as rank-1 accuracy, is

the most informative metric on an algorithm's performance, CMC(@5, 10 and 20) were also introduced due to low rank-1 accuracy values obtained when the dataset tested is very challenging. The ranking nature of the re-id problem, as opposed to traditional classification tasks, is motivated by the idea that a human operator can examine the gallery list generated and spot the correct match of the query as long as it appears within the first few candidates.

2.6.2 Mean Average Precision

When more than one ground truth is found in the gallery set with respect to a given probe, the CMC curve might not reflect the true performance of the algorithm. In fact, many algorithms could successfully retrieve an easy match at a high rank while some might miss harder positives. Therefore, the CMC in this case ignores the recall ability of the algorithm. For this purpose, the mean Average Precision (mAP) that is commonly used in information retrieval was introduced [24]. Formally, given a ranked list of gallery elements with respect to a probe, an indicator function could be used to indicate whether a gallery instance is a correct match to the probe or not as follows [160]:

$$y(i) = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ element is a correct match} \\ 0 & \text{otherwise.} \end{cases} \tag{2.18}$$

The precision at a rank k is then computed as:

$$P(k) = \frac{1}{k} \sum_{i=1}^{k} y(i). \tag{2.19}$$

Suppose there are N elements in the gallery set among which N_q ground truths correspond to query q. The average precision for query q is the mean of the precision scores taking only those corresponding to a ground truth, it is given by:

$$AP(q) = \frac{1}{N_q} \sum_{i=1}^{N} P(i)y(i), \tag{2.20}$$

where $y(i)$ is the indicator function defined in Equation 2.18. It ensures that only precision scores of ground truths are considered in the mean. Finally, the mAP score is computed over all the elements of the probe set, i.e. all the queries presented. Suppose there are S such elements, then:

$$mAP = \frac{1}{S}\sum_{q=1}^{S} AP(q). \tag{2.21}$$

In summary, for both CMC and mAP, higher scores indicate a better performance.

2.7 Summary

Person re-id is currently a very hot research topic validated by the exponential increase in the number of related publications in top computer vision venues [148]. As there are hundreds of methods that have been recently proposed, covering all these in this literature review was practically impossible. Therefore, we have attempted in this chapter to touch upon all existing categories of methods, while offering a more in-depth analysis into the works that are tightly related to this thesis. These mainly include both components of a traditional re-id system for both images and videos, namely feature extraction and matching. Although the main focus of this thesis is on video-based re-id, Chapter 3 presents an evaluation of existing image-based hand-crafted feature representations using a learnt distance metric. Hence, understanding image-based re-id in terms of person descriptors and matching methods was deemed necessary. Additionally, the video re-id task was thoroughly examined.

Chapter 3

Attribute-based Feature Evaluation

3.1 Introduction

There is no disagreement within the research community that person re-id is a very challenging problem due to significant intra-class variations caused by cross-view appearance changes including lighting conditions, viewpoint angle and pose [5, 11, 53, 54]. Despite this unanimity, no attempts were made to date to measure the amount to which each challenging attribute contributes to the performance degradation. A challenging attribute [35] designates here a certain characteristic of the images in a dataset, such as the presence of occlusions, background clutter, illumination variation, or any other property rendering the re-identification more difficult. Uncovering the individual effect of various attributes on the performance loss is crucial to identify persisting problems current systems couldn't solve, and to provide insights towards promising directions for future system design.

Feature extraction being the keystone of a person re-id system, designing features aimed at extracting the right information from person images is pivotal, especially when available datasets are small in size and thus do not contain enough instances to automatically learn these representations by the means of neural networks for instance. Consequently, many hand-crafted feature representations were proposed. They primarily rely on texture and colour cues to describe a person's appearance [7, 10–12, 54].

This work evaluates the performance of 6 state-of-the-art hand-crafted feature representations

against a collection of 4 single-shot datasets. We start by thoroughly examining the datasets in question to determine the intrinsic attributes that might affect the performance. These are found to be: illumination changes, viewpoint angle variations, occlusions, background clutter, motion blur and pose variations. We subsequently manually annotate the datasets for these attributes reflecting the different levels of difficulty each of them presents. The images are finally divided into subsets of similar properties and the evaluation of the descriptors is conducted on each subset separately. The XQDA distance metric [11] is used for matching due to its efficiency and accuracy.

The objectives of this work are as follows: (i) to identify the best-performing hand-crafted image person descriptor to date, (ii) to evaluate the robustness of existing descriptors against each of the attributes causing cross-view appearance variations, (iii) and to identify the main obstacles causing performance deterioration that current descriptors couldn't tackle successfully. This being achieved, it would guide future work on feature design.

Existing descriptors have been recently evaluated by Karanam *et al.* [53]. This included a comprehensive evaluation of 11 feature extraction schemes with 18 learned distance metrics on 16 public datasets. For each dataset, experiments of all feature-metric combinations were carried out, highlighting the best-performing combination on each benchmark. The authors briefly mention the challenges associated with each dataset, but without any indication on the most dominant attributes involved or the degree of their presence in each benchmark. Moreover, the datasets employed are of significantly different sizes which might be an important factor causing accuracy variations. For these reasons, it could not be inferred from this study which attributes have indeed caused the performance to decline, and which ones were mitigated by high-performing features.

Following the work presented in this chapter, another attempt was made towards understanding the effect of varied viewpoint angles on the re-identification performance in [161]. Highlighting the importance of quantifying the effect of visual factors on re-id systems, the authors undertake an extensive evaluation involving different viewpoint angle combinations between probe-gallery instances using 3 deeply learned models. However, to avoid the tedious annotation process of real data and to produce a controllable environment, synthetic data was generated. Although this approach is more efficient and less costly, the amount to which synthetic data is indicative of the real world cannot be easily shown, especially in terms of texture information.

Different from the work in [53], we not only determine the best-performing existing person descriptor, we rather evaluate the descriptors' robustness against each challenging case. This is vital in uncovering the reasons behind unsatisfactory performance in some cases. Revealing the remaining problems in image-based feature extraction would additionally guide video-based feature design. This lays a strong foundation for our next work on developing a spatio-temporal descriptor for video-based person re-id.

3.2 Proposed Methodology

The features evaluated, the annotation strategy adopted and the distance metric used for matching are thoroughly described in this section.

3.2.1 Features

The person descriptors selected for this evaluation are hand-crafted features that could be fed into a distance metric for matching. The most popular algorithms during the last few years with available code were employed. They include, Ensemble of Localised Features (ELF) [7], Covariance Descriptor Based on Bio-Inspired Features (gBiCov) [68], Colour Histograms and LBP Features (HistLBP) [54], Dense Colour and SIFT Features (dColorSIFT) [10], Local Maximal Occurrence (LOMO) [11], and Gaussian of Gaussian (GOG) [12].

Ensemble of Localised Features (ELF)

Along the VIPeR dataset, Gray *et al.* [7] introduced a set of low-level colour and texture features called Ensemble of Localised Features (ELF). From each person image, 29 feature channels are computed. They include 8 colour channels added to the responses of 8 Gabor [162] and 13 Schmid [163] texture filters applied to the luminance channel. The colour channels utilised are: RGB, HSV, and YCbCr noting that only one luminance channel is used, either V or Y. We recall that Gabor filter is defined as [164]:

$$g(x, y; \sigma, \gamma, \psi, \lambda, \theta) = \exp\left(-\frac{x_1^2 + \gamma^2 y_1^2}{2\sigma^2}\right) \exp\left(-i\left(\frac{2\pi x_1}{\lambda} + \psi\right)\right), \tag{3.1}$$

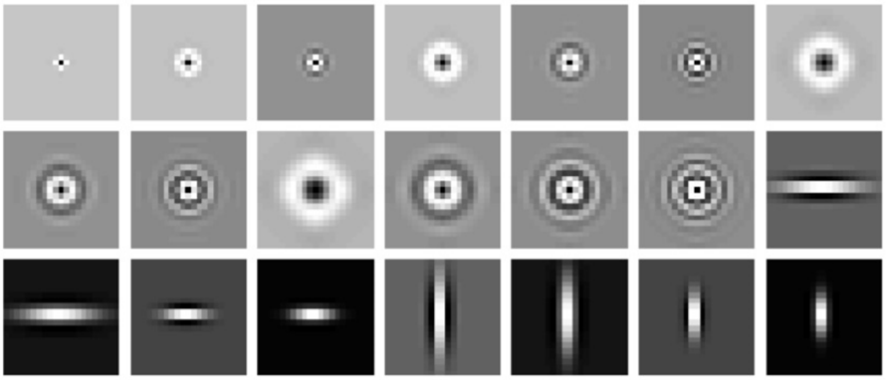


Figure 3.1: Schmid and Gabor filters used in ELF descriptor. The 13 rotationally symmetric filters are Schmid filters. The remaining are 4 horizontal and 4 vertical Gabor filters. [7]

where $x_1 = x\cos\theta + y\sin\theta$ and $y_1 = -x\sin\theta + y\cos\theta$. The remaining parameters dictate the characteristics of the filter.

Schmid filter is given by [7]:

$$F(x, y; \sigma, \tau) = \frac{1}{N}\cos\left(\frac{2\pi\tau\sqrt{x^2+y^2}}{\sigma}\right)\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right), \tag{3.2}$$

where N is a normalisation constant, and σ and τ are parameters controlling the filter's shape and characteristics. Images of the filters used in [7] can be seen in Figure 3.1. The parameters' combinations γ, ψ, λ, θ, and σ^2 for Gabor filters and τ and σ for Schmid filters used in this evaluation are kept unchanged from [7].

A part-based model is employed where images are first divided into 6 equal size horizontal stripes to approximately capture different body parts, such as the head, upper torso, lower torso, upper legs and lower legs. After feature channels computation, a 16-bin histogram is used to summarise the values in each channel of each stripe. Eventually a $16 \times 29 = 464$-dimensional vector is obtained for each stripe. The concatenation of stripes' features yields a $6 \times 464 = 2784$-dimensional feature vector for each person image.

Covariance Descriptor Based on Bio-Inspired Features (gBiCov)

Visual processing is traditionally modelled as a hierarchy of increasingly sophisticated representations as the signal propagates and reaches higher levels of the cortex [69]. A better understanding of the human visual perception inspired a number of feature representations known as

Biologically-Inspired Features (Biologically-Inspired Featuress (BIFs)) that witnessed great success in the past [69, 165, 166].

Given their impressive performance in various vision tasks including object recognition [69] and age estimation [166], Ma *et al.* advocated their use for the person re-id problem. To this end, BiCov was initially proposed [9] and was later extended to gBiCov [68]. In the latter, BIF features are extracted separately on each of the HSV colour channels using a pyramid of 24 Gabor filters taken at different scales. A max-pooling operation of corresponding pixel values is subsequently applied to the responses of each two consecutive scales which results in 12 images called in the sequel the BIF images.

To describe the information provided by BIF images, covariance descriptors are exploited. For this purpose, a 7-dimensional pixel feature vector is initially computed on each BIF image to describe its pixels' statistics. For simplicity, if B denotes a BIF image, the pixel feature vector f_i for pixel i is given by:

$$f_i = [x, y, B(x, y), B_x, B_y, B_{xx}, B_{yy}], \tag{3.3}$$

where x and y are the pixel's spatial coordinates, $B(x, y)$ is the intensity value of pixel i, B_x and B_y are the gradients in the x- and y-directions, and finally B_{xx} and B_{yy} are the second order derivatives with respect to x and y.

After computing these pixel features, each BIF image is divided into small overlapping patches of size 8×16 pixels, and covariance matrices are computed in each patch r as follows:

$$C_r = \frac{1}{n-1} \sum_{i \in r} (f_i - \mu)(f_i - \mu)^T, \tag{3.4}$$

where n is the number of pixels in patch r and μ is the mean of vectors f_i over r. The images are once more grouped in pairs and the difference of covariance descriptors for the same patch in two consecutive images is found. By consecutive images we designate here images taken at two consecutive scales.

The twelve BIF channels are also exploited separately by computing the mean magnitude in each patch, and then averaging the values obtained over two corresponding patches in two consecutive images. Subsequently, the resulting covariance descriptors and the mean magnitude of BIF fea-

tures are concatenated. This process is repeated for each of the HSV colour channels and resulting vectors are concatenated to form the final image descriptor which has 5940 dimensions.

Colour Histograms and LBP Features (HistLBP)

Local Binary Pattern (LBP) [66] features have been the cornerstone of a huge number of successful computer vision algorithms [167–172]; therefore, it was evident for researchers to attempt their use for the person re-id problem. A well-known person descriptor combining colour histograms and LBP texture features was proposed in [54].

Firstly, person images are divided into a set of overlapping regions with 50% overlap. From each region, 16-bin colour histograms are extracted from 8 colour channels similarly to ELF [7] features. Moreover, muti-scale uniform LBP histograms using 8 neighbours with a radius of 1, and 16 neighbours with a radius of 2 are computed in each patch. Channel-wise features are ℓ_1-normalised and concatenated to form the final image descriptor. The authors experimented with different region sizes yielding different number of regions and found small difference in performance by varying this parameter. Therefore, 6 regions were finally used for better efficiency. This results in an image feature size of 2580.

Dense Colour and SIFT Features (dColorSIFT)

Colour information has been a core component of nearly every descriptor in person re-id so far [7, 11, 12, 54, 68]. This is not surprising given the importance of skin, hair and clothes colours to distinguish between people. On the other hand, local feature descriptors, particularly the renowned SIFT feature [65], have seen great success in numerous vision tasks such as object recognition [65], face authentication [173], and other biometric applications [174, 175].

With the illumination and viewpoint variations problem in mind, Zhao *et al.* [10] proposed a high-dimensional person descriptor that combines both colour histograms and SIFT descriptors. It is denoted dColorSIFT. More specifically, each person image is densely divided into a set of overlapping patches of size 10×10 pixels each, taken at a step of 4 pixels. For each patch, a 32-bin colour histogram is extracted separately on the LAB colour channels. This is repeated for 3 scales of the image obtained by downsampling the original image with scale factors of 1, $3/4$ and $1/2$, respectively. All the histograms computed on the same patch are concatenated which yields

Figure 3.2: Example images from the VIPeR dataset processed with the Retinex algorithm. Images in the same column represent the same person. Left: Original images, right: Retinex processed images. [11]

a patch-wise colour feature of length $32 \times 3 \times 3 = 288$.

Furthermore, a 128-dimensional SIFT feature is computed for each patch by dividing it into 4×4 cells with 8-bin gradients histogram for each cell. As this is also performed on each colour channel separately, the resulting SIFT feature is $3 \times 4 \times 4 \times 8 = 384$-dimensional. Concatenating colour and SIFT features yields a patch-wise feature vector of 672 dimensions.

Depending on the size of the image, since dColorSIFT simply concatenates all patch-wise features after ℓ_2-normalisation, the final vector would usually be very high-dimensional. For instance, for an image size of 128×48 pixels, the final feature vector has 129,024 dimensions.

Local Maximal Occurrence (LOMO)

Local Maximal Occurrence (LOMO) features were proposed by Liao *et al.* [11] in 2015. They are specifically designed to deal with two aspects of cross-view appearance variations, illumination and viewpoint changes. Concerning the former, a pre-processing step is employed by applying the multi-scale Retinex algorithm [176] to person images before extracting the features. Retinex helps achieve colour consistency through various illumination conditions caused by lighting factors and/or camera settings. It produces images similar to the human visual perception of the scene in question. Examples of images processed by the Retinex algorithm can be seen in Figure 3.2. Scale-Invariant Local Ternary Pattern (SILTP) [70] texture features and HSV colour histograms

are subsequently extracted from the restored person images. SILTP is a variant of LBP that uses a different local comparison strategy making it more robust to intensity scale changes and image noises. Precisely, sliding windows of size 10×10 pixels with 50% overlap are used. In each sub-window, an $8 \times 8 \times 8$-bin joint histogram is extracted from the HSV colour channels. Moreover, 2 scales of SILTP texture features are computed in each patch. By examining patches in the same horizontal region, a max-pooling operation is applied to the same histogram bin across patches, thus retaining the maximal occurrence of the local pattern in each horizontal band. This is meant to deal with viewpoint angle variations that are very common in re-id.

To extract information at different scales, the original image is downsampled twice by performing a local average-pooling operation in each 2×2 neighbourhood. The same features are extracted for each scale before they are concatenated, ℓ_2-normalised, and smoothed by a log-transform to suppress large bin values. The final image feature vector has 26960 dimensions.

Gaussian of Gaussian (GOG)

Given the success of covariance and mean descriptors in many computer vision applications including person re-id, a method combining them was devised by Matsuwaka *et al.* [12] in 2016 and denoted Gaussian of Gaussian (GOG). Like many other descriptors, GOG initially divides the image into 7 overlapping horizontal regions. Each region is then divided into small overlapping patches of size 5×5 pixels with a step size of 2. To model the information presented in individual patches, a Gaussian distribution is used. To this end, each pixel i is initially described by a feature vector p_i given by:

$$p_i = [y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, R, G, B]^T, \tag{3.5}$$

where y is the y-coordinate of pixel i, M_{0° to M_{270° are the orientations along which the gradient is quantised and multiplied by the gradient magnitude, and R, G, B are the RGB colour channels. The covariance matrix Σ_H and the mean μ_H of pixel feature vectors in each patch are computed and summarised using a Gaussian distribution as follows:

$$\mathcal{N}(p; \mu_H, \Sigma_H) = \frac{\exp\left(-\frac{1}{2}(p-\mu_H)^T \Sigma_H^{-1}(p-\mu_H)\right)}{(2\pi)^{d/2}|\Sigma_H|}, \tag{3.6}$$

where $|\cdot|$ is the matrix determinant operator and d is the dimension of pixel feature vector p.

Patches belonging to the same horizontal stripe are weighted according to their distance from the central vertical axis of the image using a bell-shaped function to account for background clutter. They are subsequently summarised using another Gaussian distribution. Modelling a horizontal region by a unique Gaussian is meant to deal with viewpoint angle variations.

Each patch or region Gaussian is flattened into the tangent Euclidean space upon its computation using a matrix logarithm. This step is essential to obtain a final feature that falls in the Euclidean space and is hence suitable for use with off-shelf distance metrics. The final image representation is the concatenation of all stripes' features. It has 7567 dimensions.

3.2.2 Annotations

There is currently no objective measure to quantify the complexity of re-id datasets, and developing such a measure is not within the scope of this thesis. Alternatively, we are interested in assessing how challenging attributes, as perceived by humans, can individually affect the performance of re-id algorithms. For this purpose, we undertake a careful examination of available re-id datasets to investigate pertinent attributes characterising cross-view appearance changes. After identifying 6 such attributes, we manually annotate 4 popular public single-shot datasets for the identified challenging attributes reflecting various levels of complication they add to the re-identification process. The attributes considered and the annotations carried out are as follows.

Viewpoint Angle

There are realistically two reasons causing two instances of the same person to be captured with different angles. Firstly, surveillance cameras are installed at different positions with various elevations and angles depending on the topology of the monitored area. Secondly, the direction pedestrians are moving with respect to a given camera modifies the viewing angle. Consequently, for instance, a top view of a person might need to be matched with a horizontal view, or a front view might need to be matched with a side view.

Since the datasets employed in this evaluation present mainly a horizontal view, we annotate the images for the viewpoint angle attribute in a similar manner to Gray *et al.* annotations of the VIPeR dataset in [7]. More specifically, the viewpoint angle label for each person image is obtained by estimating the angle formed by the optical axis of the camera and the normal to the

Figure 3.3: Example images on the viewpoint angle label. The labels from left to right are: $0°$, $45°$, $90°$, $135°$, $180°$, $225°$, $270°$, and $315°$.

person's body. Computing an exact value for such angle is not easy and in fact not needed. Alternatively, the estimated angles are quantised into 8 bins: $0°$, $45°$, $90°$, $135°$, $180°$, $225°$, $270°$ and $315°$. For instance, a front view of the person corresponds to $0°$, a side-right view to $90°$, a back view to $180°$, a side-left view to $270°$, and so forth. Example images with their viewpoint angle labels can be seen in Figure 3.3. To summarise a pair of matched images with one label for this attribute, we compute the absolute value of the difference between the viewpoint angles of the images concerned. We subsequently obtain a label characterising the viewpoint angle difference between each positive pair.

Illumination Changes

Different camera settings and lighting conditions, such as the nature of the light source illuminating the scene, can cause images of the same person to appear considerably different across camera views. A change in illumination produces apparent colour change in corresponding images, thus affecting the performance of image processing techniques that use colour information in their design [177]. Since most descriptors in person re-id rely on the holistic appearance to distinguish between individuals, colour is therein a key feature. Consequently, assessing the robustness of available descriptors against varied illumination is essential.

It is possible to use a pixel-wise difference method to measure the level of illumination variation between two images of the same scene taken under different conditions. However, this is not possible when the images represent two different scenes. For this reason, we resort to manually analysing the different levels of illumination existing in the chosen datasets by visually comparing images of the same person. Subsequently, the varying degree of illumination between corresponding images is estimated over a scale ranging from 0 to 3, 0 being the lowest and 3 the highest.

Figure 3.4: Example images on the illumination variation label. Two adjacent images represent a positive pair. Two example pairs are chosen for each illumination variation degree. The labels from left to right are: 0, 1, 2 and 3, indicating increasing degree of illumination change.

These values reflect the apparent change in colours in the two corresponding images. Some examples can be seen in Figure 3.4.

Occlusions

The occlusions problem is faced by most surveillance applications such as action recognition, anomaly detection, and person re-id for the following reasons. The areas usually monitored by surveillance cameras represent critical infrastructure of a city such as airports, train stations, and shopping centres. These places are kept under surveillance because they usually exhibit a higher risk of being targeted by terrorist or criminal activity. Since these sites are vital to the urban life, they are often crowded. This causes a person's trajectory to be occluded multiple times before leaving the camera view. Other than self-occlusions between people, occlusions with objects present in the scene could occur. Concerning the re-id problem, significant occlusions will contribute into larger intra-class variations by confusing a person's identity with another subject or object that might have a totally different appearance.

To assess the effect of this attribute on the performance, the datasets selected for this evaluation exhibit various levels of occlusions, they are annotated as follows. The fraction of the body of the occluded person is estimated in each image, and values are quantised into four different levels

Figure 3.5: Example images on the occlusions labels. Grouped images represent the same occlusions degree. The labels from left to right are: 0, 1, 2 and 3, indicating increasing degree of occlusions.

from 0 to 3, 0 being no occlusions involved and 3 a high degree of occlusions. Some examples are illustrated in Figure 3.5. Since individual labels are obtained for each image, a matching pair's label is computed by adding image-wise labels, thus approximating the total degree of occlusions exhibited by each corresponding pair of images.

Background Clutter

The presence of background clutter could play a major role in increasing the complexity of person re-id. Although persons' bounding boxes are provided in a traditional re-id scenario, addressing this issue is still crucial. The features are usually extracted from the whole image, and two images for different people with a similar background might exhibit higher similarity than two images of the same person with different cluttered backgrounds, even when the person occupies the most pixels in the image. To resolve this issue, one might suggest performing background subtraction or person segmentation. However, these tasks are far from trivial especially for the type of data used in re-id that is typically collected with low quality cameras, and may suffer from frequent and significant occlusions.

Quantifying the background clutter attribute is practically not feasible. Therefore, we use binary annotations indicating the presence or absence of background clutter in an image with labels 1 and 0, respectively. Examples are shown in Figure 3.6. Finally, the amount of background clutter involved in a matched pair of images is found by adding the labels of corresponding images.

Figure 3.6: Example images on the background clutter labels. Grouped images have the same label. The group to the left has a label of 0 indicating the absence of background clutter, and the group to the right has a label of 1 indicating the existence of background clutter.

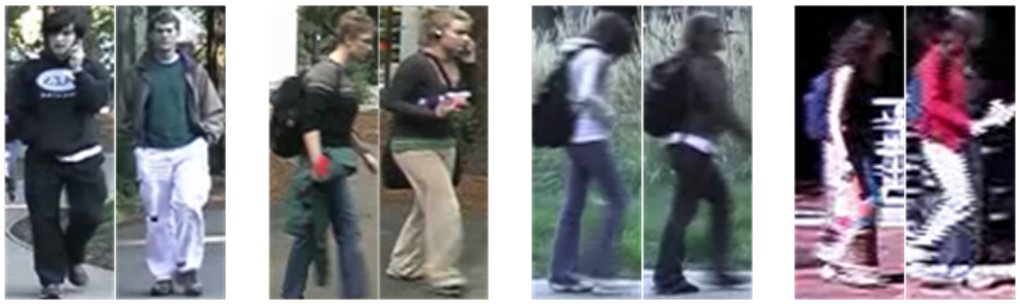


Figure 3.7: Example images on the motion blur labels. Grouped images have the same label. The labels from left to right are: 0, 1, 2, and 3, indicating increasing levels of motion blur.

Motion Blur

One of the common artefacts in images is motion blur. It mainly occurs when the camera is moving or the subject being filmed is moving. It causes the latter to appear smeared along the direction of the relative motion. Motion blur is generally more visible in images or video frames captured by cameras having a low shutter speed, thus allowing the scene being recorded to change within a single exposure. Since the presence of motion blur was noticed in re-id datasets, we evaluate the effect of this attribute on the performance.

To annotate the images, the level of motion blur was determined based on the degree of its visual strength in the image. A scale of 4 levels was initially used from 0 to 3, 0 being the lack of blur and 3 a significant blur level. It was then noticed that only a few instances present a high blur level which renders the experimentation with these impossible. Therefore, these labels were then shrunk into two only by merging 0 and 1 into 0, and 2 and 3 into 1. Examples can be seen in Figure 3.7. Similarly to occlusions and background clutter attributes, image-wise labels of corresponding images are added indicating the total level of blur observed in a positive pair of images.

Figure 3.8: Example images on the pose variations labels. Each two adjacent images are a correct match. The two pairs on the left have a label of 0 indicating no pose variation, and the two pairs on the right have a label of 1 indicating the existence of pose variations.

Pose Variations

Although people are typically walking in a re-id scenario, there might be cases where they stop or walk with varying paces causing them to be captured with different poses. In addition, even when they are simply walking, the deformable nature of the human body exhibits various postures depending on the phase of the walking cycle they are at when captured, for instance whether it's a stance phase or a swing phase.

Since walking is the only activity involved in the datasets used in this evaluation, we denote by pose variations the phase of the walking cycle the person is at in a given image, and the order of their feet in that phase. A label of 1 is given when the person is at the same phase with the same feet order in corresponding images, and 0 otherwise. See Figure 3.8 for examples.

Finally, six labels are obtained for each pair of corresponding images describing the aforementioned challenging attributes. This enables an in depth analysis of the causes inflicting performance degradation of existing person descriptors.

3.2.3 Learned Distance Metric

The popularity of distance metric learning in re-id [11, 54, 85, 108, 143, 154] emerged from its ability to considerably boost the performance at a low computational cost. A renowned learned distance metric is Cross-View Quadratic Discriminant Analysis (XQDA). Liao *et al.* [11] introduced XQDA to complement their LOMO features achieving remarkable results. Unlike its closest counterpart KISSME [85] that uses PCA for dimension reduction before learning the distance matrix, XQDA learns a subspace W and a distance function simultaneously through the following

equation [11]:

$$d_W(x,z) = (x-z)^T W(\Sigma_I^{'-1} - \Sigma_E^{'-1})W^T(x-z), \tag{3.7}$$

where d_W is the distance between feature vectors x and z, and W is the subspace projection matrix. If Σ_I and Σ_E are the respective covariance matrices of the intrapersonal (positive samples) and extrapersonal (negative samples) differences, then $\Sigma'_I = W^T\Sigma_I W$ and $\Sigma'_E = W^T\Sigma_E W$.

The variances σ_I and σ_E are then used to separate between the two classes in the projected subspace. This could be achieved for each projection direction w by maximising $\frac{\sigma_E(w)}{\sigma_I(w)}$. Since $\sigma_E(w) = w^T\Sigma_E w$ and $\sigma_I(w) = w^T\Sigma_I w$, the column vectors of matrix $W = (w_1, ..., w_r)$ are computed by maximising the Generalised Rayleigh Quotient:

$$J(w) = \frac{w^T\Sigma_E w}{w^T\Sigma_I w}, \tag{3.8}$$

which is equivalent to

$$\max_w w^T\Sigma_E w, \quad s.t. \quad w^T\Sigma_I w = 1. \tag{3.9}$$

This optimisation problem has a closed-form solution obtained by eigenvalue decomposition, that is, by computing the eigenvectors of $\Sigma_I^{-1}\Sigma_E$. The eigenvectors corresponding to eigenvalues larger than 1 are collected to form the columns of the learned subspace $W = (w_1, ..., w_r)$.

3.3 Experiments

3.3.1 Datasets

VIPeR

The release of the Viewpoint Invariant Pedestrian Recognition (VIPeR) dataset [7] is one of the reasons that triggered increasing interest in person re-id, and led to the development of a plethora of methods. This can be largely attributed to it being very challenging in terms of low-resolution images characterised by significant illumination and viewpoint angle variations, in addition to background clutter and motion blur. Furthermore, VIPeR is a single-shot dataset where each person is represented by only a pair of images to be matched, thus lacking enough visual and temporal information to overcome cross-view appearance changes. Technically, VIPeR includes

Figure 3.9: First ten image pairs from the VIPeR dataset. Images in the same column represent the same person taken from two different views.

1,264 images of 632 pedestrians taken from two outdoor camera views over the course of several months in an academic setting. Bounding boxes are manually drawn around persons, and the resulting images are scaled to 128×48 pixels. In addition to persons' identities, Gray *et al.* [7] provide viewpoint angle annotations. Examples images selected from VIPeR can be seen in Figure 3.9.

GRID

The underGround Re-Identification (GRID) dataset [44], also known as QMUL GRID, was released by the Queen's Mary University of London's computer vision research group in 2009. It involves 250 pairs of pedestrian images captured by 8 disjoint cameras in a busy underground station. Moreover, the gallery set includes 775 distractors, i.e. person images that have no correct match in the probe set. They are added to increase the size of the gallery set in order to further complicate the re-identification task. These are ignored in some benchmarking results, and also in this study as matched pairs are needed to describe cross-view appearance changes. Similarly to VIPeR, the images in GRID are manually cropped and annotated for person identities. They are of a fairly low spatial resolution with various crop sizes, and they exhibit high levels of illumination and viewpoint changes, background clutter, and occlusions. Example images from GRID can be seen in Figure 3.10.

Figure 3.10: First ten image pairs from the GRID dataset. Images in the same column represent the same person taken from two different views.

PRID450S

Person Re-ID 450S (PRID450S) [178] is a single-shot dataset that was released in 2014 based on the Person Re-ID 2011 (PRID2011) video dataset. It comprises 450 image pairs of persons captured by two static surveillance cameras in an outdoor environment. Person images are manually cropped and annotated without performing any scaling into a fixed size similarly to VIPeR. Moreover, the authors provide automatically generated motion-based human silhouettes that are further manually segmented into parts including the head, torso, and legs with labels indicating potential carried objects. The images in PRID450S exhibit significant levels of illumination and viewpoint angle variations but with limited background clutter or occlusions. In addition, there are noticeable clothes similarities between people which adds more complexity to people matching. Selected images from PRID450S can be seen in Figure 3.11.

i-LIDS

The imagery Library of Intelligent Detection Systems (i-LIDS) is a benchmark for video analytics covering various scenarios such as event detection and tracking. It was initially released by the UK's government for the purpose of research in 2013. From the Multiple-Camera Tracking Scenario (MCTS) of this benchmark, a person re-id dataset was created by Wang *et al.* [33]. The dataset released covers two re-id protocols: single-shot and video-based re-id. We employ the single-shot version in this work and its video-based counterpart (iLIDS-VID) in the chapters to follow. i-LIDS includes 600 images of 300 pedestrians captured by two non-overlapping surveil-

Figure 3.11: First ten image pairs from the PRID450S dataset. Images in the same column represent the same person taken from two different views.

Figure 3.12: First ten image pairs from i-LIDS dataset. Images in the same column represent the same person taken from two different views.

lance cameras at an airport arrival hall. The persons are manually cropped and annotated, and the bounding boxes obtained are of size 128×64 pixels. i-LIDS bears varied lighting conditions and viewpoint angles in addition to background clutter, occlusions, and clothing similarities, thus making it very challenging. Example images from i-LIDS can be seen in Figure 3.12.

Measuring the effect of each challenging attribute on the features' performance necessitates the evaluation to be conducted on each label separately. This requires a sufficient number of examples per label to learn the distance metric and obtain stable results. To meet this requirement, the four datasets were merged into a single set containing 1,632 image pairs. This was necessary since none of the benchmarks covers all the difficulty levels for all the attributes simultaneously.

3.3.2 Evaluation Protocols

A classic person re-id evaluation protocol involves only two camera views, the probe and the gallery. Motivated by the release of some datasets that include more than two camera views such as Market1501 [24] and DukeMTMC-reID [133], a more realistic scenario has been investigated lately. It aims to build a system that is able to retrieve any instances relevant to the query in all other available disjoint views. We use a similar setting here as the dataset employed in the experiments is obtained by merging four different datasets collected in various places with significant spatial and temporal separation. It is both important and interesting to evaluate the features' performance in similar more realistic conditions that could outline their generalisation abilities. In reality, it is very hard to collect enough data from concerned camera views, i.e. the probe person view and the target gallery view, and have them readily annotated before re-identification.

For the experimental work, two evaluation protocols are employed. In either case, we first proceed into extracting the six types of features from all the images in the merged datasets. Each challenging attribute is then considered separately, and the dataset is divided into subsets representing each individual label for this attribute. For instance, if the illumination variation case is considered, four subsets are obtained each representing an illumination variation degree (from 0 to 3). For each probe image presented in a subset, matching using XQDA distance metric is conducted, the ranks of the correct matches are recorded, and the results are reported.

Regarding the first protocol, the evaluation is conducted on all the images in a subset, regardless of the differences among subsets' sizes. More specifically, for the illumination variation attribute for instance, the evaluation is carried out on all the data in each illumination degree subset although the size of the subset for a label of 3 is 76 image pairs while for a label of 1 it is 678 pairs. We refer to this protocol as the Unbalanced Evaluation.

It has been noticed that differences in the number of examples are significant for some labels. Therefore, in order to alleviate any effect the size of the subset might have on the system's performance, the evaluation was repeated ensuring the same number of examples for each case. For example, if one illumination variation degree presents 76 instances and the others present more instances, only 76 instances are randomly selected from each subset. This is referred to as Balanced Evaluation.

To ensure stability, 100 runs of each algorithm are performed for each experiment by dividing the subset randomly into half for training and half for testing. The average results in top-1 matching rate (rank-1 accuracy) are reported. Note that when the number of examples in a category is small, learning XQDA metric was not possible. Hence, results are not reported.

It is worth noting that the original code provided by the authors is employed in our experiments, except for the ELF algorithm where an implementation by Layne *et al.* [75] is used. The default parameter settings of the feature descriptors, such as the patch size, the number of horizontal stripes, and any other method specific parameters are kept unchanged as they are deemed the most convenient by the primary authors.

A more comprehensive evaluation could be performed by considering two or more attribute combinations and assessing the effect of the increased level of difficulty they provide. However, this was not possible as there are not enough available examples for various labels' combinations of two or more attributes.

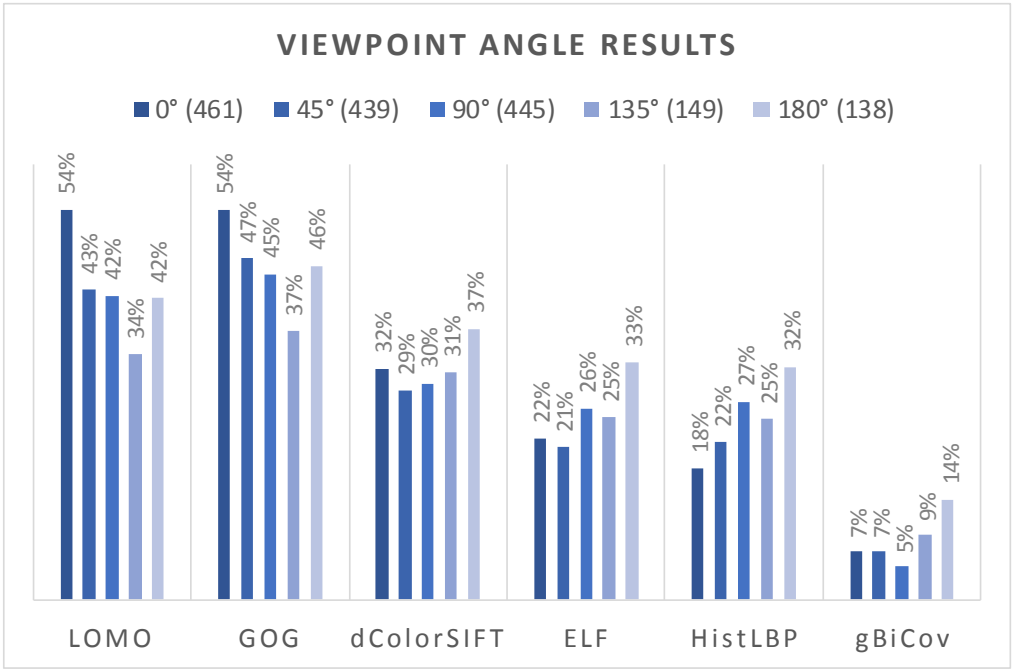
3.4 Results and Discussion

The results for both unbalanced and balanced evaluations in rank-1 accuracy (top-1 matching rate) are shown in Figure 3.13 and Figure 3.14, respectively. The analysis on each type of evaluation is conducted separately.

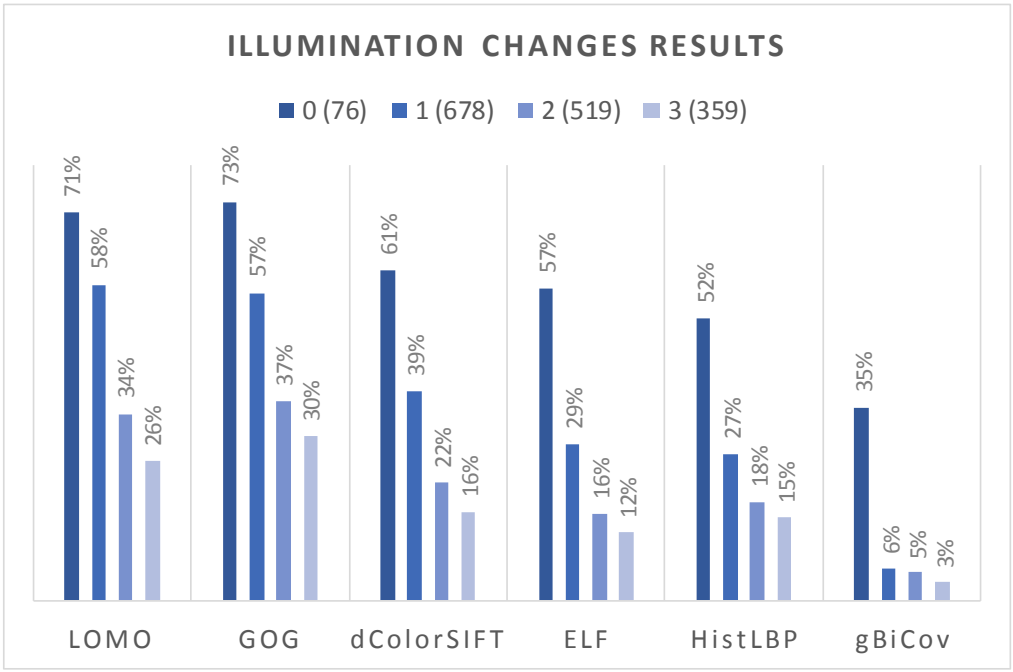
3.4.1 Unbalanced Evaluation

From the results obtained on the unbalanced evaluation (Figure 3.13), the following conclusions can be made:

- GOG and LOMO are by far the best performing features with LOMO showing better ability to scale to large datasets while GOG showing superiority in dealing with challenging cases. The performance of GOG declines significantly compared to LOMO with large datasets, such as the pose variations subset with label 1 that includes 1264 examples, or the occlusions subset with label 0 containing 1340 examples. When the dataset is relatively small, GOG exhibits better performance for more challenging cases, such as bigger viewpoint angle difference and higher levels of illumination changes and occlusions. LOMO explicitly

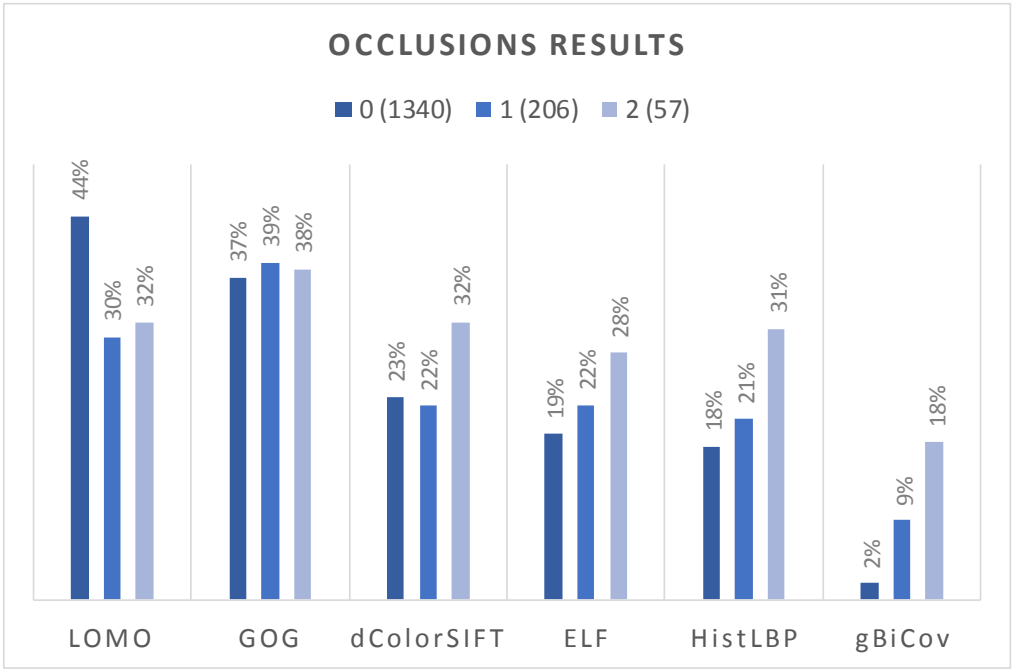


(a)

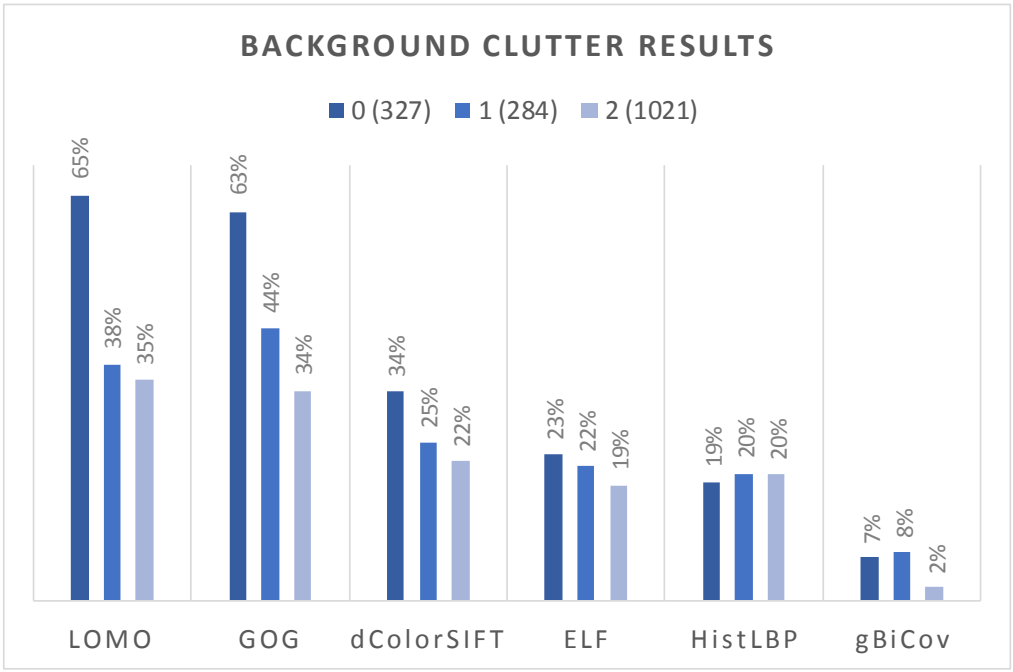


(b)

Figure 3.13: Unbalanced evaluation results in top-1 matching rate. The numbers in parentheses represent the number of image pairs in the label's subset.

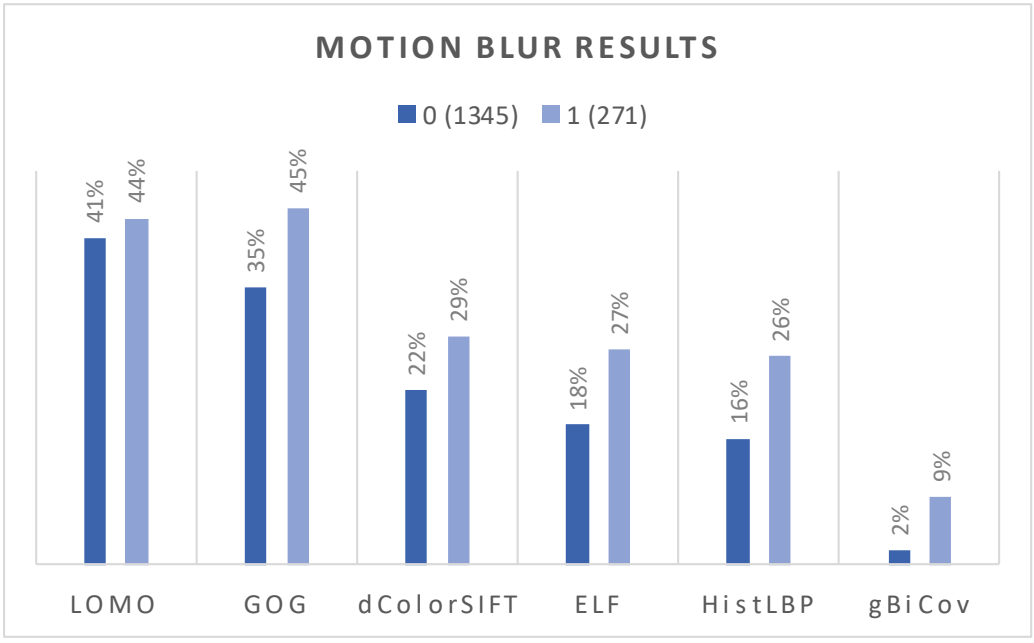


(c)

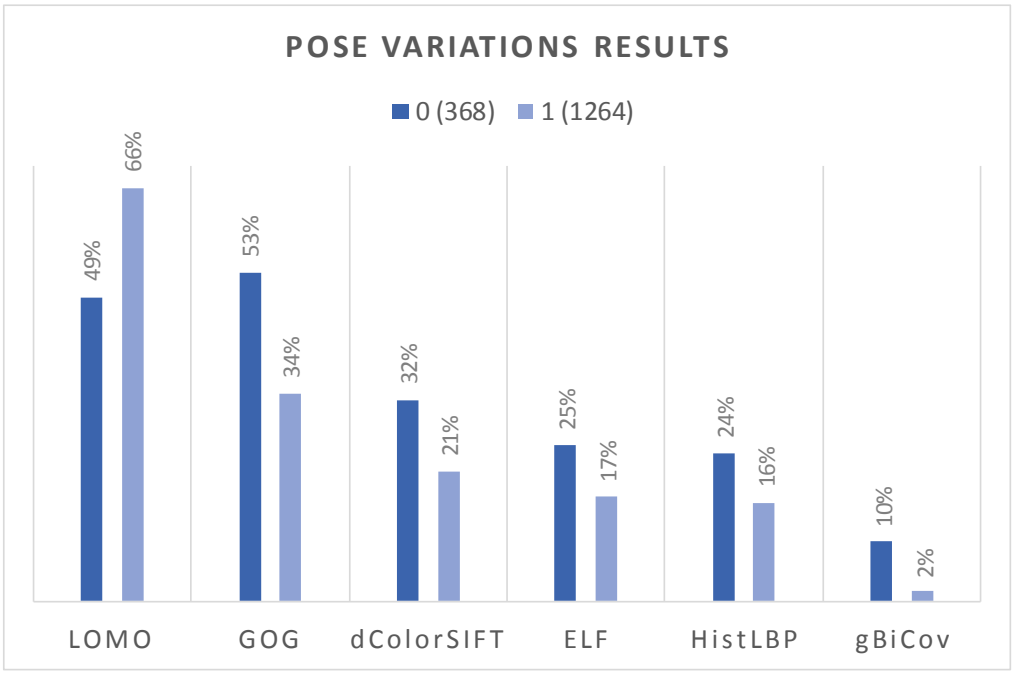


(d)

Figure 3.13: Unbalanced evaluation results in top-1 matching rate. The numbers in parentheses represent the number of image pairs in the label's subset.

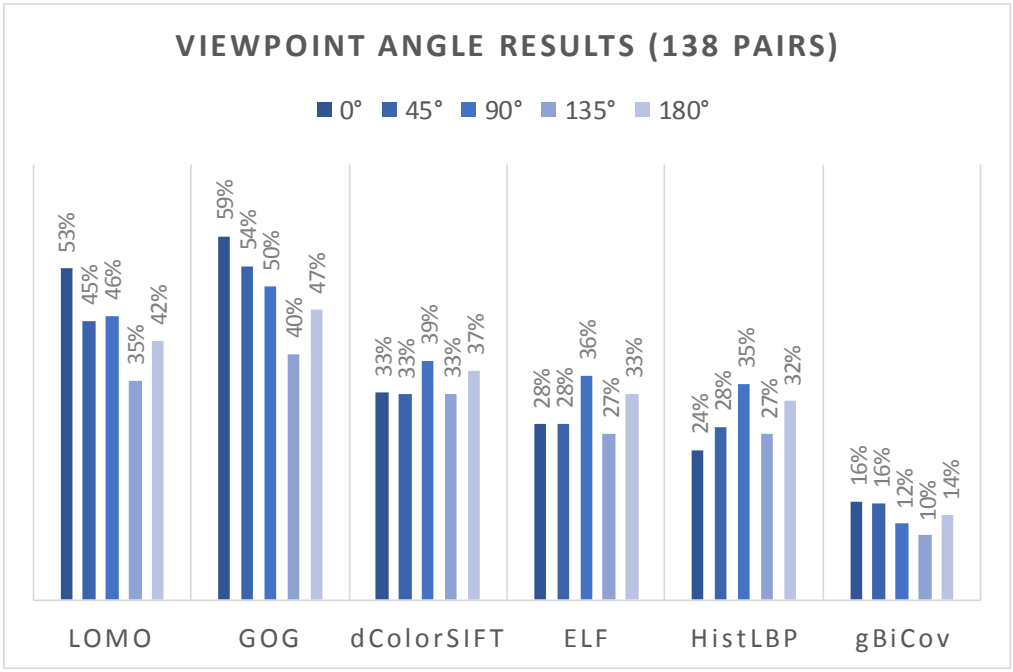


(e)

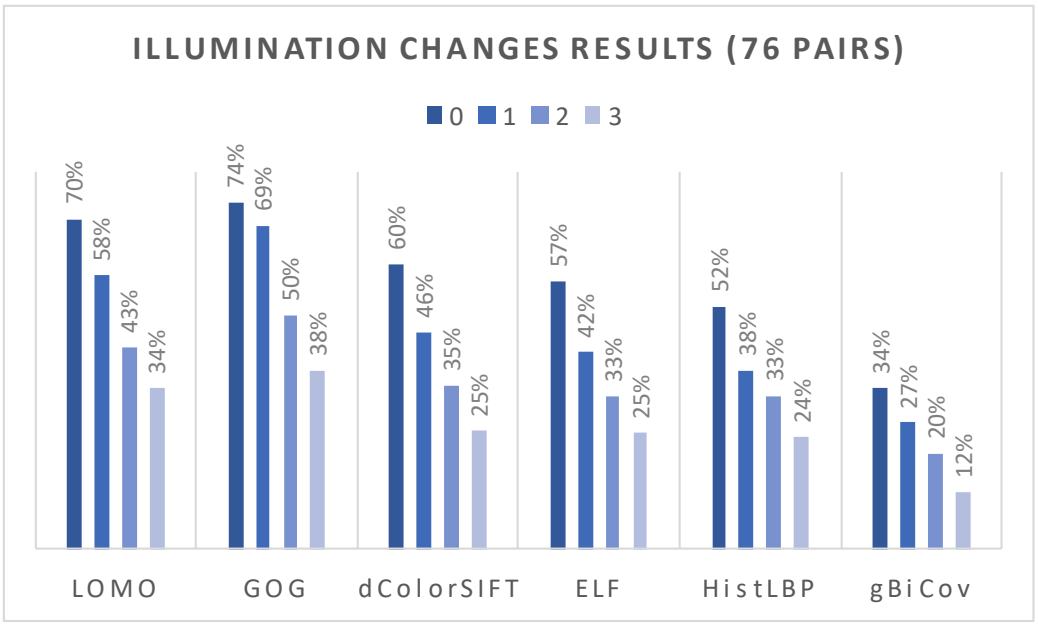


(f)

Figure 3.13: Unbalanced evaluation results in top-1 matching rate. The numbers in parentheses represent the number of image pairs in the label's subset.

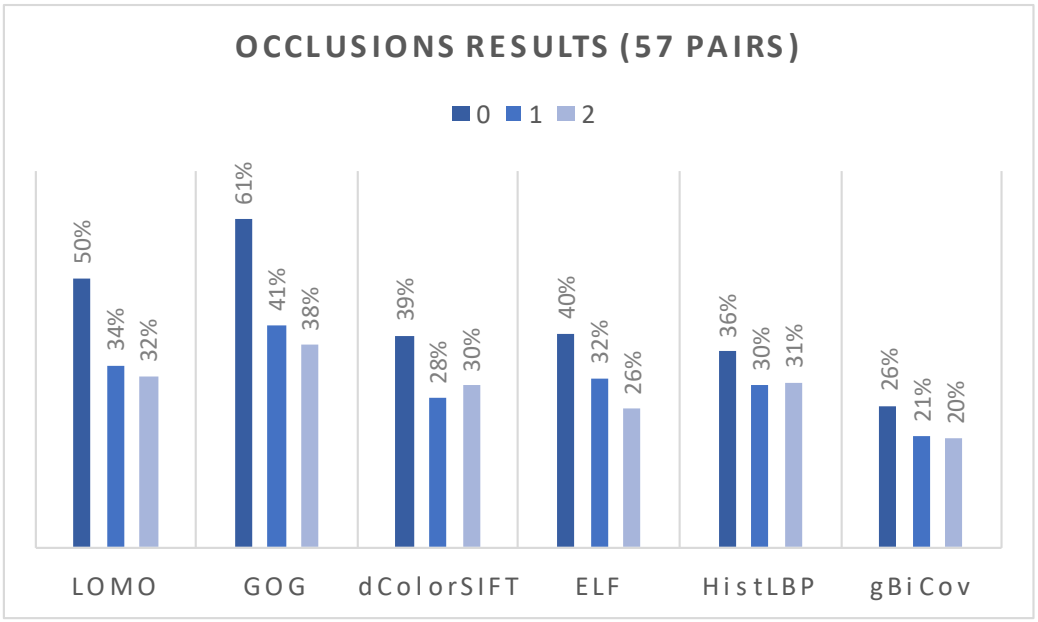


(a)

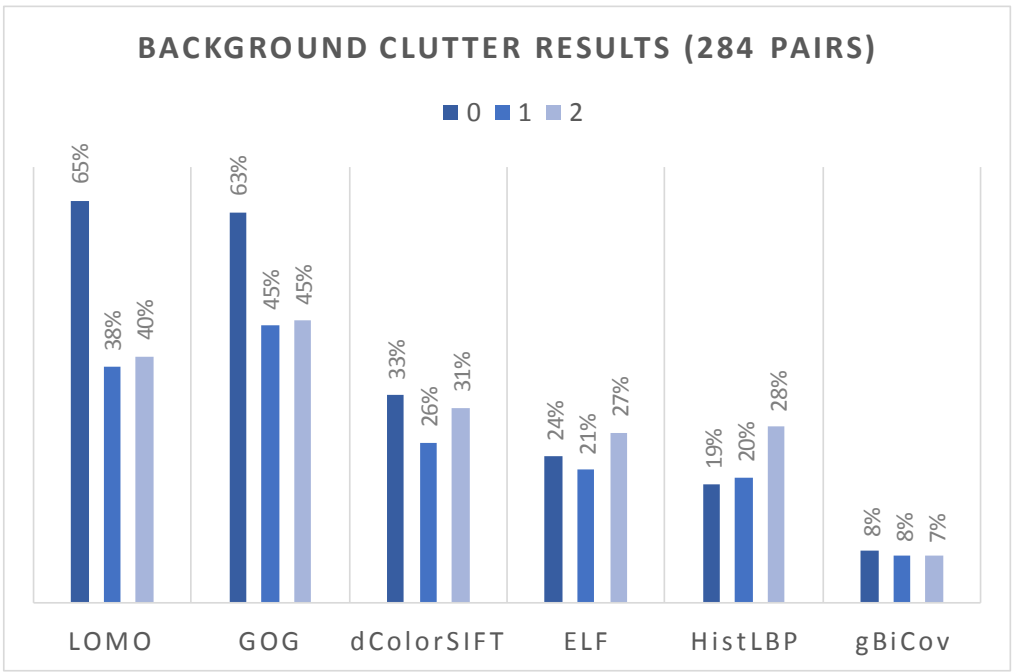


(b)

Figure 3.14: Balanced evaluation results in top-1 matching rate.

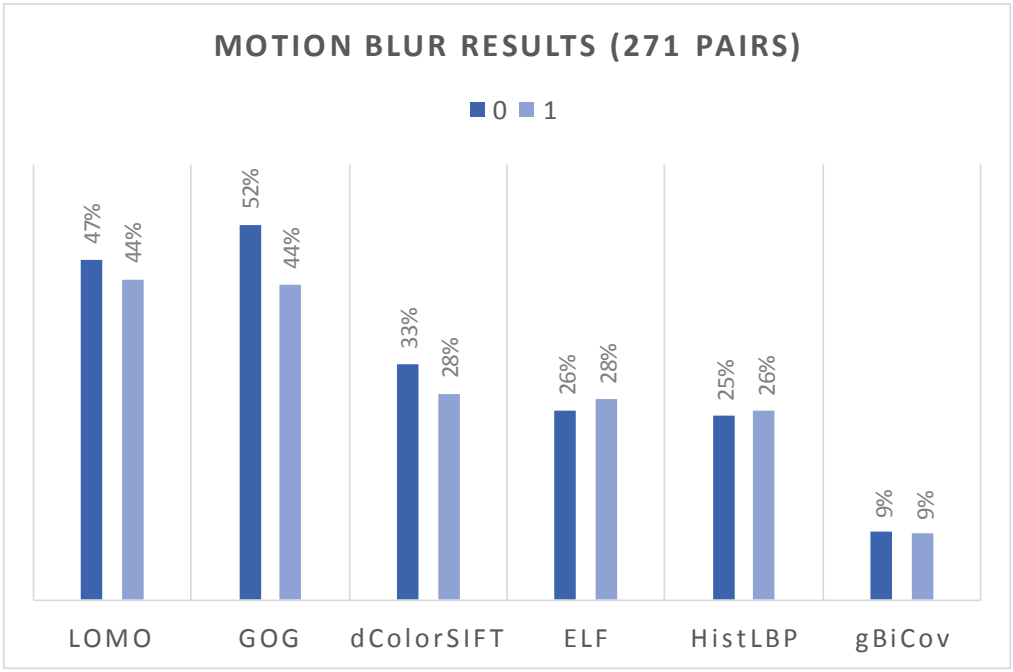


(c)

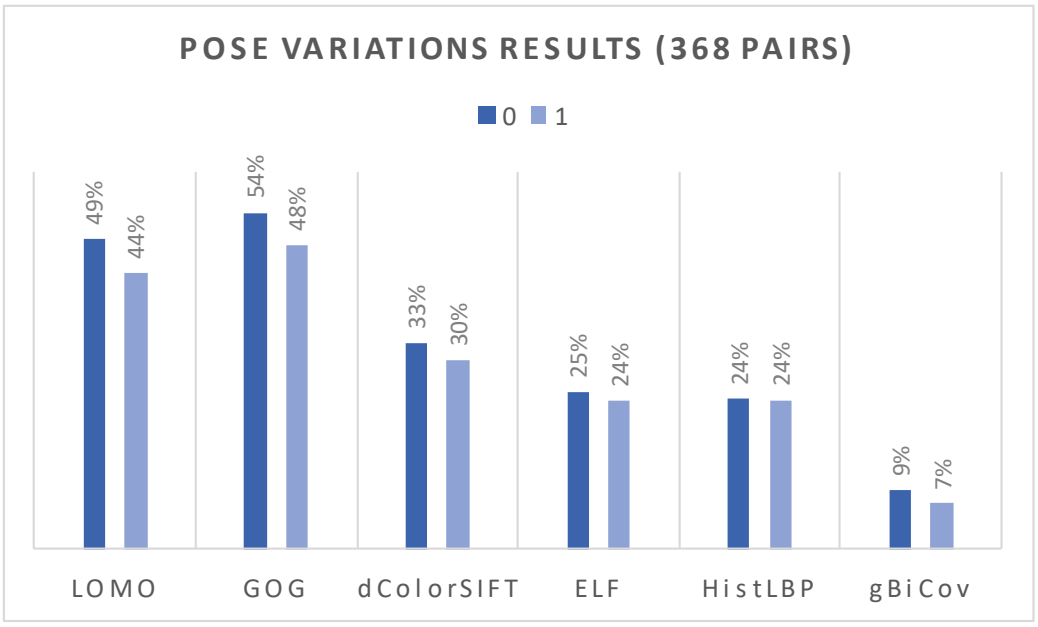


(d)

Figure 3.14: Balanced evaluation results in top-1 matching rate.



(e)



(f)

Figure 3.14: Balanced evaluation results in top-1 matching rate.

uses a pre-processing technique to mitigate illumination changes, and employs horizontal local maximal pooling to deal with viewpoint angle variations. On the other hand, GOG specifically tackles background clutter by patch weighting, and viewpoint angle and illumination changes through hierarchical modelling of colour and texture pixel features using Gaussian distributions. These elaborate designs prove more successful compared to the other descriptors.

- Another highlight of the results obtained from the unbalanced evaluation is that illumination variation consistently inflicts the most performance deterioration regardless of the features and the number of examples used. The decrease in accuracy reaches a maximum of 45% for LOMO and a minimum of 32% for gBiCov.

- For the viewpoint angle attribute, LOMO and GOG achieve the best rank-1 accuracy of 54% for 0° difference, i.e. same viewpoint angle for probe and gallery images. Meanwhile, the remaining methods produce their best performance for a label of 180°, possibly due to this label's subset being the smallest in size.

- Regarding the remaining attributes: occlusions, background clutter, motion blur and pose variations, one of the labels in each case had a significantly larger subset (over 1000 pairs) compared to the others (less than 400 pairs). In all the cases except LOMO for occlusions and GOG for pose variations, the biggest subset suffered from the lowest performance. The exceptions show the ability of LOMO to better deal with larger datasets compared to the other methods, and the low influence of pose variations on the performance.

Further conclusions regarding the effect of datasets' challenging attributes on the performance could not be made from this evaluation. This is due to the significant difference in the size of the subsets obtained for each case. As this could be a factor influencing the performance, we focus our analysis on the balanced evaluation where the same subset size is used for each challenging case. This would better infer the attribute's contribution to the performance rather than the dataset's size.

3.4.2 Balanced Evaluation

From the results obtained on the balanced evaluation, the following conclusions can be made:

- As Figure 3.14 shows, the high-performing GOG and LOMO features exhibit almost similar performance patterns, i.e. they perform well or under-perform under almost the same conditions. However, the overall performance of GOG is slightly better than LOMO. This further implies that even the most successful feature representations still struggle to deal with certain challenging cases.

- Illumination changes have proved again to contribute the most into performance degradation. The accuracy drop when the degree of illumination rises between positive matches is considerable for all six features. It is at least 24% for gBiCov and rises up to 36% for GOG and LOMO.

- Motion blur and pose variations have a very small effect on the performance for all studied features as can be seen in the last two graphs of Figure 3.14. Since colour and texture are the most prominent cues relied upon by these descriptors, they are hardly affected by motion blur. The latter usually impacts the image sharpness that doesn't play a big role in re-id. Moreover, poses do not vary considerably in the datasets employed here as people are mainly walking.

- Although the same viewpoint (angles difference of 0°) yields a better accuracy for GOG and LOMO compared to different viewpoints, and an angle difference of 135° yields the lowest performance for 5 out of 6 features, viewpoint angle difference is not a major player in causing accuracy decline. This is possibly due to a large focus in dealing with viewpoint angle variations within existing feature representations that proved partially successful.

- A high degree of occlusions or the presence of background clutter seems to have a considerable effect on the feature's performance as the decline is nearly between 18% and 25% in rank-1 accuracy for each attribute for both GOG and LOMO. On the other hand, a small difference in accuracy is observed between subsets labelled 1 and 2 in both cases. As a label of 1 indicates that only one image of the matched pair involves the attribute (occlusions or background clutter) and a label of 2 often refers to both images in a pair having this characteristic, re-identification in either case is made equally harder.

3.4.3 Key Findings and Insights

The findings of this study also provide the following directions for possible investigation.

- Image-based and particularly single-shot person re-id is challenging when the images to be matched exhibit significant appearance changes. Given the wide availability of video sequences from surveillance cameras, it is intuitive to make use of the rich visual and temporal cues from person tracks to circumvent these challenges.
- It was shown that GOG descriptor yields the most promising performance. Therefore, it would be useful to build upon it to develop a robust person descriptor for video-based person re-id.
- Despite superior performance, GOG descriptor still presents some drawbacks that need to be further addressed. Colour and texture cues are apparently not enough to deal with illumination variations. Therefore additional information should be incorporated, and the most suitable colour channels (less sensitive to illumination conditions) should be investigated.
- Apart from lighting conditions, the largest performance drop for GOG is observed in the case of severe occlusions. When leveraging GOG in video-based re-id, a method should be devised to reduce the presence of noisy occluded frames in video sequences, or to automatically select the most suitable frames for matching.

3.5 Summary

The intrinsic challenges associated with a typical person re-id scenario causing important appearance changes across disjoint camera views were identified in the study detailed within this chapter. A data annotation strategy to quantify the level of difficulty these challenging attributes bring to the re-identification process was also developed. An evaluation was subsequently conducted using six renowned hand-crafted image-based feature representations on a collection of four fully annotated single-shot datasets using XQDA distance metric for matching.

The main findings of this work can be summarised as follows. Firstly, GOG is currently the best-performing hand-crafted image person descriptor. Nonetheless, it is not sufficiently robust to high levels of dataset complexity. Secondly, among the six attributes studied, illumination variations

have the most impact on reducing the accuracy of existing person descriptors, including GOG. Therefore, this attribute should be dealt with carefully in future feature design.

In the next chapter, we build on the main conclusions drawn here by extending GOG into 3 dimensions to tackle the video-based re-id task, while addressing its drawbacks. In addition to video re-id being the most realistic surveillance scenario, leveraging all types of information provided by person videos is essential to engineer feature representations that are robust to cross-view appearance variations.

Chapter 4

3D Gaussian Descriptor

4.1 Introduction

The dramatic effects cross-view appearance variations inflict on the re-id performance have been extensively demonstrated in the previous chapter. Despite noticeable improvement over the years, even the most advanced feature representations for image-based re-id fail to deal with some challenging attributes, particularly illumination changes. Given the insufficient amount of information provided by single-shot images, overcoming these hurdles is not simple. However, luckily the single-shot scenario is not the most realistic one. Traditionally, video segments are available for persons of interest from different camera views in a CCTV network. Therefore, ignoring the rich visual cues and the potentially useful temporal information that is easily accessible from videos is not the best option. Motivated by this fact, we explore video-based person re-id in this chapter.

Video-based re-id has witnessed an ample amount of research in the past few years [14, 31, 138, 142, 147, 155, 179, 180]. However, not enough work has been invested in developing hand-crafted feature extraction schemes that are specifically tailored for this protocol. Alternatively, most of the descriptors used leverage low-level image-based feature representations describing spatial information while completely ignoring any temporal cues. To compensate for using relatively simple representations, adequate matching methods such as metric learning or multi-shot ranking based on set-to-set distance measures [13, 14, 154] or frame selection and weighting [137, 138] were developed. Despite the matching method being an important element of a re-id system, its per-

formance can be largely affected by the input features. This was empirically proved in Chapter 3 where a large gap in accuracy was observed with different feature representations, despite using the same distance metric XQDA. In addition to a wider focus on the matching method, developing deep models to automatically learn the feature representations drew a great deal of attention lately [31, 116, 142, 144, 181–184]. These methods proved extremely successful in some scenarios, however they require a large amount of data for model training, and thus still suffer with small datasets.

A number of hand-crafted spatio-temporal descriptors exist in the literature. The most notable ColHOG3D [33] and STFV3D [34] are described in Section 2.4.1. The main drawback of these descriptors is that they require performing walking cycle extraction on person tracklets before computing the features. The poor quality of the videos captured by surveillance cameras, and the repeated occlusions and clutter throughout a person's trajectory render this task far from trivial. Another downside of these algorithms is that they do not put enough effort into mitigating the effect of cross-view appearance changes including different viewpoint angles and illumination degrees. This oversight substantially affects their performance, and thus their ability to compete with other learning-based methods.

Motivated by the lack of a robust hand-crafted person descriptor that fully leverages both spatial and temporal cues, and that can be efficiently computed and used with common distance metrics, a feature extraction scheme for video-based person re-id is proposed in this chapter. There have been very successful attempts in the past to extend 2-dimensional features into 3 dimensions, such as HOG3D [150] and SIFT3D [185]. Moreover, to benefit from the advances achieved in image-based person re-id feature design and from the results obtained in the attribute-based feature evaluation conducted in Chapter 3, the high-performing GOG feature is improved and extended into 3 dimensions yielding GOG3D (3-Dimensional Gaussian of Gaussian). The proposed descriptor is further combined with the best existing supervised metric learning methods to enhance its accuracy. Details of the proposed algorithm and the experiments carried out to validate its performance are thoroughly described in this chapter.

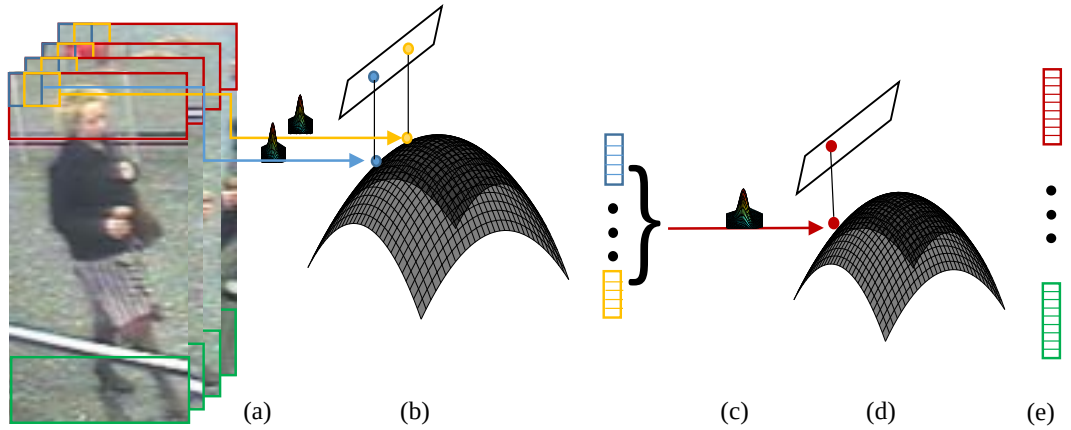


Figure 4.1: Diagram representing GOG3D feature extraction scheme. (a) Patch Gaussians are computed. (b) They are then flattened into the Euclidean space. (c) Region Gaussians are formed using patches in the same horizontal region. (d) They are also projected into the Euclidean space. (e) Region feature vectors are concatenated to form the final image feature. Finally, average-pooling is performed over image features of the same person sequence.

4.2 Proposed 3-Dimensional Gaussian of Gaussian (GOG3D)

The proposed GOG3D descriptor is first briefly described in this section before examining closely each aspect of the algorithm. Like the vast majority of hand-crafted re-id features including GOG [12], a part-based model is adopted where each person image is divided into R overlapping horizontal stripes, roughly representing different human body-parts. Local information in each region R is subsequently described using small overlapping patches. Pixel information is used to model each patch by a multivariate Gaussian distribution. The patch Gaussians obtained are then embedded into the space of Symmetric Positive Definite (SPD) matrices which allows their projection into the Euclidean space. To account for viewpoint angle variations, patches belonging to the same horizontal region are in turn summarised by a Gaussian that is flattened into the Euclidean space forming the region feature vector. Region feature vectors are then concatenated to form the final image feature. To obtain the final representation of a person sequence, the features of the images constituting the sequence are average-pooled. A diagram summarising GOG3D is shown in Figure 4.1.

4.2.1 Pixel Features

The backbone of GOG3D is the pixel feature vector that depicts the type of information most suited for re-id. This is also where the main modification lies compared to GOG. Considering an image (frame) from the person sequence, a 10-dimensional feature vector is used to describe each of the pixels. It summarises the spatial position of the pixel, its spatial and temporal gradient information, and the intensity values of some colour channels. Namely, for a pixel i, the pixel feature vector p_i is given by:

$$p_i = [x, y, M_{0^\circ}, M_{90^\circ}, M_{180^\circ}, M_{270^\circ}, |I_t|, L, A, B]^T, \tag{4.1}$$

where x and y are the x- and y-coordinates of pixel i taken from the top-left of the image, M_{0° through M_{270° are the orientations along which the gradient is quantised and multiplied by its magnitude, $|I_t|$ is the gradient magnitude in the temporal direction, and finally L, A, B correspond to the Lab colour channels. Each dimension of p_i is then scaled to the range [0, 1]. The computation of these pixel features is detailed in the following.

Spatial Coordinates

Assuming that person images are spatially aligned, adding the pixel's location to its feature vector is essential in preserving the local structure inside each region. This can be achieved using the x- and y-coordinates. Ideally, these should be computed separately from the top-left of each region. However, since regions are overlapping, the same pixel might have different coordinates in different regions which increases the computational complexity. This could be avoided by setting the coordinates from the top-left of the image without causing any problems since the images are spatially aligned.

In the GOG algorithm, only the y-coordinate is used. This was motivated by the observation that person images are aligned vertically but not horizontally [186]. However, as it will be detailed in the sequel, we find that the orderless representation of patches in the same horizontal stripe as a means to deal with viewpoint angle variations can cause the loss of some important spatial information that is useful for re-identification. This can be avoided by including the horizontal location of the pixel represented by its x-coordinate.

Spatial Gradients

Let $S = \{Q_k | k = 1, ..., N\}$ be a person sequence of images such that N is the number of frames in this sequence. By taking a pixel $i(x, y, t)$ in frame Q_t, the gradients I_x, I_y and I_t in the horizontal, vertical and temporal directions can be computed as:

$$I_x(x, y, t) = i(x + 1, y, t) - i(x - 1, y, t), \tag{4.2}$$

$$I_y(x, y, t) = i(x, y + 1, t) - i(x, y - 1, t), \tag{4.3}$$

$$I_t(x, y, t) = i(x, y, t + 1) - i(x, y, t - 1), \tag{4.4}$$

where x, y and t are the x-, y- and t-coordinates of pixel $i(x, y, t)$, respectively.

Similarly to GOG, we leverage the spatial gradients I_x and I_y to describe texture and shape cues. To this end, the orientation O and magnitude M of vector $G_i = \begin{pmatrix} I_x \\ I_y \end{pmatrix}$ taken at pixel i are computed as:

$$O = \arctan(I_y / I_x), \tag{4.5}$$

and

$$M = \sqrt{I_x^2 + I_y^2}. \tag{4.6}$$

Although it is possible to use these values directly in the pixel feature vector p_i, this is not believed to be the best option [187]. In fact, it has been argued in [187] that quantisation into vector angles rather than using magnitude and orientation raw values is essential to enhance the discriminative power of the descriptor. Therefore, the values of O are quantised into bins. Two methods are known for this task: hard voting and soft voting. In the former, the bins' boundaries are defined and all the weight is assigned to the only bin containing the value in question, the remaining weights are all set to 0. Alternatively, in soft voting, weights are assigned into two neighbouring bins according to the distance of the concerned value from each bin boundary. This will be formally explained in

the next paragraph. Compared to hard voting, soft voting reduces the loss of information caused by quantisation.

Four bins are considered here as a compromise between the vector's dimension and its discriminative power. Since the orientation O is a value between 0° and 360°, the reference points or possible bin boundaries are 0°, 90°, 180°, and 270°. Let α and β be the boundaries of the bin containing O, that is, $\alpha \leq O < \beta$ where $\alpha, \beta \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, the distances (positive differences) $d_\alpha = |O - \alpha|$ and $d_\beta = |O - \beta|$ from O to the bin boundaries are computed, and the voting weights are assigned as $w_\alpha = d_\beta/(d_\alpha + d_\beta)$, $w_\beta = d_\alpha/(d_\alpha + d_\beta)$ and $w_{(\theta \neq \alpha,\beta)} = 0$. To emphasise larger gradients, these weights are multiplied by the gradient magnitude M to obtain $M_\theta, \theta \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$.

Temporal Gradient

Encoding the temporal correlation between consecutive frames of a video sequence is a powerful way to describe the motion information involved. Given the significant illumination, viewpoint, and pose variations suffered in a typical person re-id application, colour and texture cues alone are not enough to match cross-view instances of a person. For this purpose, temporal information is exploited here by the means of the temporal gradient. More specifically, the magnitude of I_t computed in Equation 4.4 is found by taking its absolute value $|I_t|$ and adding it to the pixel feature vector.

Although, it is possible to compute the gradient orientation in 3 dimensions and quantise it using a regular polyhedron in a manner similar to HOG3D in spirit [150], this is suboptimal in our case due to the following reasons. Firstly, binning in 3D while preserving the distinctive power of the descriptor requires the use of a dodecahedron (12 bins) or icosahedron (20 bins) [150], thus raising significantly the dimension of the pixel feature vector p_i. A high-dimensional p_i will cause numerical problems upon the computation of covariance matrices in small patches. On the other hand, separating the temporal gradient from the spatial ones is triggered by the idea that spatial gradients serve as a texture/shape descriptor while motion information is encoded via the temporal gradient. Some examples of the information obtained by computing $|I_t|$ for all sequence frames and taking their average can be seen in Fig. 4.2. This highlights the type of information added by computing the magnitude of the gradient in the temporal direction.

Figure 4.2: The information encoded by the temporal gradient. The top row represents frames sampled from person sequences of iLIDS-VID dataset, each 2 adjacent frames are taken from corresponding sequences. The bottom row represents the temporal gradient computed for these sequences and average-pooled over each sequence's frames.

Colour Channels

The choice of the colour channels is crucial for any appearance descriptor, especially for person re-id where individuals are mainly distinguished by their clothes' colours and texture. Hence, selecting the best colour channels for the task at hand is of great importance. In order to exploit the advantages brought by various colour spaces, some researchers resorted into extracting the features a number of times, each with different colour channels such as RGB, HSV, Lab and normalised RGB (nRGB), before concatenating them like in $\text{GOG}_{\text{fusion}}$ [12] and mean of Moments (moM) [74] descriptors. Despite enhancing the performance, this practice is highly inefficient as the features have to be extracted four times for each person, and the resulting vector has a high dimension which in turn slows down the matching process. For this purpose, it is convenient to select the most robust colour channels that can better cope with illumination changes.

Unlike GOG that uses the RGB colour channels, we opt to use the Lab (CIELAB) colour space in GOG3D. The advantages brought by the Lab channels over their RGB counterparts can be summarised as follows. Firstly, Lab components are designed to mimic the human visual perception, that is, a certain change in L, a, and b values roughly corresponds to a similar amount of change perceived by humans. In addition, Lab channels are non-linear and device-independent with respect to a reference white point, which is more aligned with the non-linearity of the human visual

perception in colours [188]. The L, a, and b components can be briefly described as follows. L represents lightness and ranges from 0 to 100, a and b channels are for green-red and blue-yellow, respectively. They usually range from -100 to 100 with negative values representing green/blue and positive values representing red/yellow for a and b channels, respectively. An example frame from iLIDS-VID dataset with the ten pixel feature channels computed is shown in Figure 4.3.

4.2.2 Patch Gaussians

As both mean and covariance features have proved successful in person re-id [33, 68], a promising way to leverage both types of information is to summarise them using a Gaussian distribution. As discussed in [12], it is definitely possible to use a GMM instead for a more accurate representation. However, given the small patch size, a simple Gaussian model should be sufficient to describe the information presented in a patch. More importantly, unimodal Gaussians can be efficiently projected into the Euclidean space which renders the matching process with the resulting feature much easier, as any off-the-shelf distance metric can thus be exploited. Therefore, using pixel feature vectors in each patch H, the mean μ_H and covariance Σ_H are computed as:

$$\mu_H = \frac{1}{n_H} \sum_{i \in H} p_i, \tag{4.7}$$

and

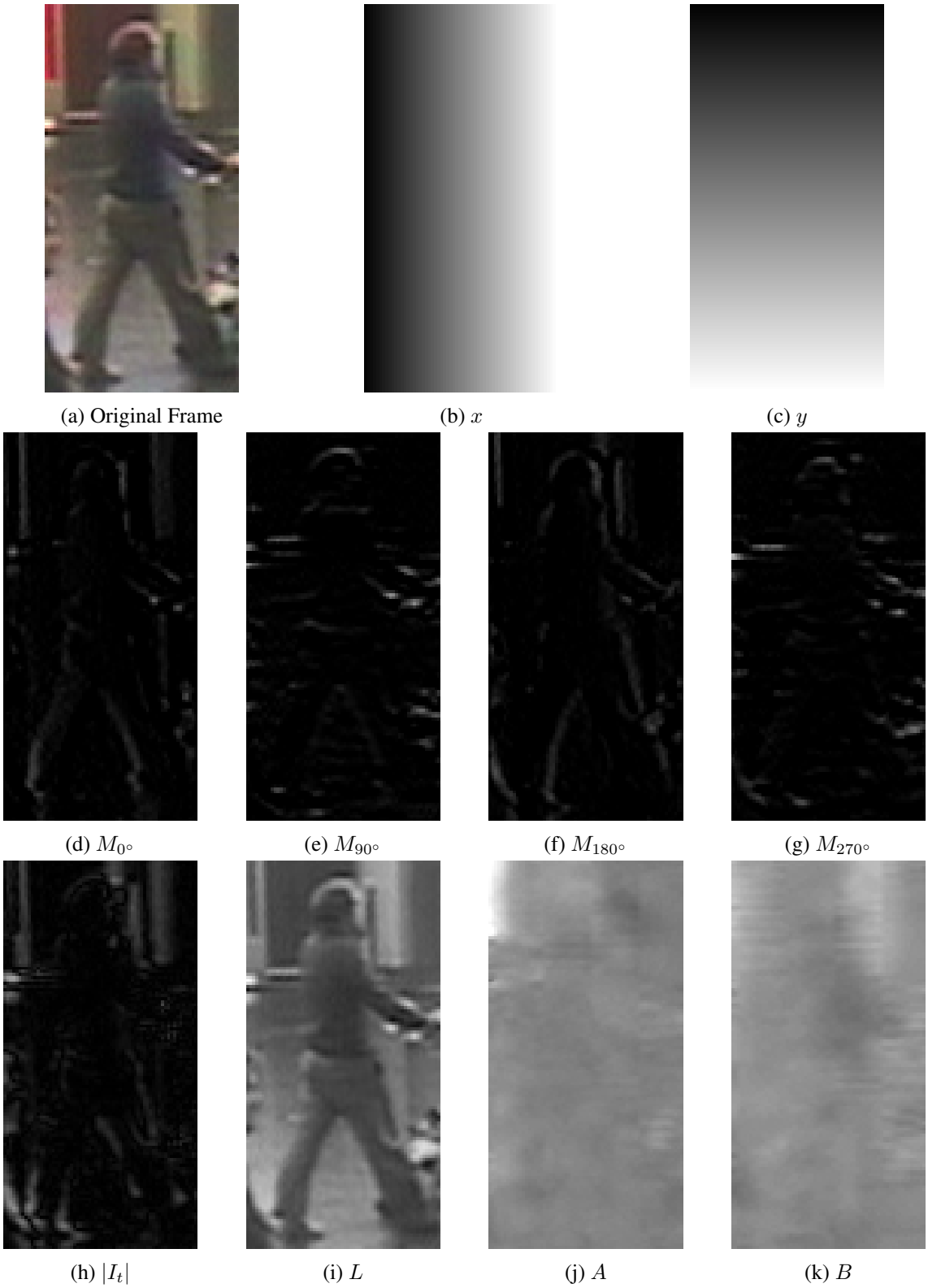
$$\Sigma_H = \frac{1}{n_H - 1} \sum_{i \in H} (p_i - \mu_H)(p_i - \mu_H)^T, \tag{4.8}$$

where n_H is the number of pixels in patch H and p_i is the feature vector of pixel i defined in Equation 4.1. Subsequently, the patch Gaussian $\mathcal{N}(p; \mu_H, \Sigma_H)$ is given by:

$$\mathcal{N}(p; \mu_H, \Sigma_H) = \frac{\exp\left(-\frac{1}{2}(p - \mu_H)^T \Sigma_H^{-1} (p - \mu_H)\right)}{(2\pi)^{d/2} |\Sigma_H|}, \tag{4.9}$$

where $|\cdot|$ is the matrix determinant operator and d is the dimension of pixel feature vector p.

Once all patch Gaussians are computed, an algorithm that will be explained in Section 4.2.4 is used to project these Gaussians into the Euclidean space transforming them into patch feature vectors f_H.



(a) Original Frame (b) x (c) y

(d) M_{0° (e) M_{90° (f) M_{180° (g) M_{270°

(h) $|I_t|$ (i) L (j) A (k) B

Figure 4.3: Random frame taken from iLIDS-VID dataset with the corresponding 10 pixel feature channels. The values are scaled to the range [0,1]. In the feature channels ((b) to (k)), the lighter the pixel the higher the value.

Figure 4.4: The shape of the function dictating the patch weights. More weight is assigned to the patches closer to the central vertical axis of the image where the person is centred to account for background clutter.

To account for background clutter, patches are weighted similarly to SDALF algorithm [189] according to their distance from the central vertical axis of the image. A patch weight w_H is given by:

$$w_H = \exp\left(-\frac{(x_H - x_C)^2}{2\sigma^2}\right), \tag{4.10}$$

where $x_C = W/2$, $\sigma = W/4$, x_H is the x-coordinate of the central pixel in patch H, and W is the image width. Based on this definition, it is easy to see that more weight is assigned to patches closer to the central vertical axis of the image where the person is expected to be centred as can be seen in Figure 4.4.

4.2.3 Region Gaussians

In a similar manner to patch Gaussian computation, the patches belonging to the same horizontal region are in turn summarised into a region Gaussian using their mean and covariance information. Based on the patch weights found in Equation 4.10, the region mean μ_R and covariance Σ_R are defined for region R as:

$$\mu_R = \frac{1}{\sum_{H \in R} w_H} \sum_{H \in R} w_H f_h, \tag{4.11}$$

and

$$\Sigma_R = \frac{1}{\sum_{H \in R} w_H} \sum_{H \in R} w_H (f_H - \mu_R)(f_H - \mu_R)^T, \tag{4.12}$$

where f_H is the patch feature vector for patch H obtained by projecting patch Gaussians into the Euclidean space. The region Gaussians are consequently computed according to Equation 4.9 and projected into the Euclidean space before concatenation to form the final representation of an image.

To obtain the video sequence representation, frame-wise feature vectors are average-pooled over the person's sequence with the aim of reducing the noise and enriching the representation. It is also worth noting that covariance matrices are regularised by adding a small value ϵ to their diagonal entries in order to prevent them from becoming singular. Namely, for a matrix Σ, the regularised matrix Σ' is given by $\Sigma' = \Sigma + \epsilon I$ where I is the corresponding identity matrix. Following [12], ϵ slightly differs for a patch compared to a region covariance matrix. For a region R, ϵ is the product of a small number ϵ_0 and the trace norm of matrix Σ_R, i.e. $\epsilon = \epsilon_0 \operatorname{Tr}(\Sigma_R)$. In patches that do not include a lot of variability in their pixel values, $\operatorname{Tr}(\Sigma_H)$ approaches zero. Therefore, for a patch H, ϵ is set as $\epsilon = \epsilon_0 \max(\operatorname{Tr}(\Sigma_H), 10^{-2})$.

4.2.4 Euclidean Space Projection

Projecting patch and region Gaussians into the Euclidean space is essential for GOG3D, primarily to obtain a final descriptor that can be used with off-the-shelf distance metrics that are often designed to be applied to features in the Euclidean space. This can be achieved in a two-stage process as follows.

It is well known that multivariate Gaussian distributions lie on a Riemannian manifold that can be embedded into the space of SPD matrices [190]. Such embedding is favoured as the SPD space endowed with the log-Euclidean metric can be locally flattened into the tangent Euclidean space through matrix logarithm. More specifically, consider a d-dimensional multivariate Gaussian $\mathcal{N}(\mu_H, \Sigma_H)$, this can be embedded into a $(d+1)$-dimensional SPD matrix P_H as follows:

$$\mathcal{N}(p; \mu_H, \Sigma_H) \sim P_H = |\Sigma_H|^{-\frac{1}{d+1}} \begin{bmatrix} \Sigma_H + \mu_H \mu_H^T & \mu_H \\ \mu_H^T & 1 \end{bmatrix}. \tag{4.13}$$

P_H can subsequently be mapped into the Euclidean tangent space by computing $\Gamma_H = \log(P_H)$ where $\log(\cdot)$ is the matrix logarithm. Noting that Γ_H is a symmetric matrix, only the upper triangu-

lar part needs to be stored resulting in the final vector f_H being $m = (d^2+3d)/2+1$ dimensional. This is done through half-vectorisation as follows:

$$f_H = \text{vec}(\Gamma_H) = [\Gamma_H(1,1), \sqrt{2}\Gamma_H(1,2), ..., \sqrt{2}\Gamma_H(1,d+1), \Gamma_H(2,2), \sqrt{2}\Gamma_H(2,3), ..., \Gamma_H(d+1,d+1)]^T, \quad (4.14)$$

where $\Gamma_H(i,j)$ is the (i,j) element of matrix Γ_H. Off-diagonal entries in this equation are multiplied by $\sqrt{2}$ to ensure that the Frobenius norm of Γ_H remains equal to the ℓ_2-norm of f_H, that is $||\Gamma_H||_F = ||f_H||_2$. We recall that the Frobenius norm of a matrix $A = (a_{ij}) \in \mathbb{R}^{m\times n}$ is given by $||A||_F = \sqrt{\Sigma_{i=1}^{m}\Sigma_{j=1}^{n}|a_{ij}|^2}$ and the ℓ_2-norm of a vector $x = (x_1, ..., x_n)$ is given by $||x||_2 = \sqrt{x_1^2 + ... + x_n^2}$.

4.2.5 Mean Removal and ℓ_2-Normalisation

The importance of normalisation techniques especially for high-dimensional feature vectors has been extensively demonstrated in previous literature [191–195]. This procedure is particularly crucial when different types of features are fused together such as the colour and gradient attributes employed in GOG3D (see Figure 4.3). Otherwise, features with a bigger numerical range would dominate those with a smaller range [194], and the distance measure would be controlled by biased dimensions.

To rectify the effect of different dimensions having different numerical ranges and statistical distributions, similarly to [12], mean removal and ℓ_2-normalisation are applied to the final features computed. If f is a given feature vector and $\overline{f}$ is the mean of all feature vectors in the training set, the normalised feature vector f_N is given by:

$$f_N = \frac{(f - \overline{f})}{||f - \overline{f}||_2}. \quad (4.15)$$

In the normalised feature space, the features are mean-centred, i.e. their mean is 0, and each vector has an ℓ_2-norm of 1.

4.2.6 Implementation Details

The 2-dimensional GOG algorithm uses 7 overlapping horizontal regions (50% overlap) with a patch size of 5×5 pixels. A step size of 2 pixels is employed to extract the patches, and the parameter ϵ_0 is set to $\epsilon_0 = 0.001$. For a fair comparison when directly comparing to GOG, a similar setting is used for GOG3D. This results in the patch feature vector being of size $(10^2 + 3 \times 10)/2 + 1 = 66$, and the region feature vector being $(66^2 + 3 \times 66)/2 + 1 = 2,278$ dimensional, finally the person descriptor obtained is $7 \times 2,278 = 15,945$ dimensional.

When comparing against state of the art (Section 4.3.5), an optimal setting is used for GOG3D. It involves more horizontal regions, a bigger patch size of 9×9 pixels, and a smaller $\epsilon_0 = 0.0001$. The patch extraction interval is kept at 2 pixels. More horizontal stripes contribute into a finer body part decomposition. Bigger patches deal favourably with potential misalignment in a horizontal region and allow for a smaller ϵ_0 to be used by increasing the number of data points compared to the feature dimension. Results using 10 and 15 horizontal regions are presented. They are denoted GOG3D^{10} and GOG3D^{15}, respectively. In this case, the resulting feature vector is 22,780 and 34,170 dimensional for $R = 10$ and $R = 15$, respectively. The use of more than 15 horizontal stripes didn't bring additional performance gain while raising the computational complexity by increasing the feature dimension.

After mean removal and ℓ_2-normalisation are applied to GOG3D features, dimension reduction using PCA is performed with KISSME metric [85]. The dimension of the reduced feature is set to 100 similarly to [53]. A linear kernel is used with kNFST [108], kLFDA [54], and kMFA [54] in all experiments. The same setting used for GOG3D is also adopted for GOG, that is, frame-wise features are averaged for a person sequence. As per the common evaluation protocol, the datasets are randomly partitioned into half for training and half for testing. This process is repeated 10 times and the average results in CMC top-matching rates are reported.

4.3 Experiments

4.3.1 Datasets

Similarly to other related spatio-temporal descriptors, GOG3D is evaluated on the two most widely tested benchmarks for video-based person re-id, PRID2011 [51] and iLIDS-VID [33].

PRID2011 [51] is a video benchmark corresponding to PRID450S single-shot dataset. It involves two adjacent static surveillance cameras A and B capturing video sequences of 385 and 749 persons, respectively. Among these, 200 individuals appear in both camera views. The lengths of the sequences vary from 5 to 675 frames averaging at almost 100 frames per sequence. Persons are manually cropped and annotated for this dataset, and the size of the crops provided is 64×128 pixels. PRID2011 is considered challenging mainly because it involves significant illumination changes and viewpoint angle variations. A common evaluation protocol for PRID2011 is to consider only the matched sequences (200 pairs) and discard the remaining ones [33, 34, 147, 155, 179, 180, 196]. This makes it a closed-set re-id scenario with no distractors. Example video sequences from PRID2011 can be seen in Figure 4.5.

iLIDS-VID [33] dataset involves 300 identities with 2 sequences per person resulting in 600 videos. Sequences lengths range from 23 to 192 frames with an average number of 73 frames per sequence. iLIDS-VID is the video-based version of iLIDS dataset used in Chapter 3. It is taken at an airport arrival hall using two disjoint camera views. Similarly to PRID2011, persons are manually cropped and annotated. All the crops have equal size of 64×128 pixels. This dataset is very challenging due to significant cross-view illumination and viewpoint angle variations, occlusions, and background clutter. Example sequences from iLIDS-VID are shown in Figure 4.6.

4.3.2 Matching Methods

After frame-wise GOG3D features are extracted from person sequences and average-pooled over the frames constituting each sequence, a feature vector is obtained for each person tracklet. This in fact renders the matching process similar to single-shot re-id. Consequently, given a query instance, an adequate matching method should be employed to retrieve its correct match. For this purpose, the kNFST supervised subspace learning method is used. Moreover, to highlight the consistent performance gain of GOG3D compared to GOG regardless of the matching method employed, an evaluation using KISSME [85], kLFDA [54], kMFA [54] and XQDA [11] algorithms is conducted. More details on the theoretical background of these algorithms can be seen in Section 2.3.3.

Figure 4.5: Example sequences from PRID2011 dataset. The first ten frames of each sequence are shown. Adjacent rows represent a correct match (same person from different views).

Figure 4.6: Example sequences from iLIDS-VID dataset. The first ten frames of each sequence are shown. Adjacent rows represent a correct match (same person from different views).

Dataset	**iLIDS-VID**			**PRID2011**		
Rank	1	5	20	1	5	20
GOG + XQDA	66.6	87.2	96.9	86.6	97.6	99.6
GOG_I_t + XQDA	70.6	90.6	98.0	87.2	97.2	99.8
GOG_I_t_x + XQDA	72.9	91.3	**98.7**	87.7	97.8	99.9
GOG3D + XQDA	**73.7**	**92.0**	98.3	**89.9**	**97.9**	**100**

Table 4.1: Components analysis of GOG3D. Best results in top-matching rates are in bold.

4.3.3 Components Analysis

To evaluate the effect of the modifications made to the pixel feature vector on the performance, an experimental evaluation is conducted by progressively adding different pixel features to baseline GOG. We recall that the vector used in GOG [12] is given in Equation 3.5. The $|I_t|$ component is first added to that vector, and the descriptor obtained is denoted GOG_I_t. The x-coordinate is subsequently added yielding GOG_I_t_x. Finally, GOG3D is obtained by replacing the RGB colour channels by the Lab channels. In order to accurately analyse the effect of these changes, the parameters are kept unchanged from GOG as detailed in Section 4.2.6. Moreover, the XQDA matching method employed with GOG in [12] is equally used. The results obtained in top-matching rates are shown in Table 4.1.

It can be clearly observed from these results that the modifications made to the pixel feature vector bring a noticeable improvement in accuracy especially on iLIDS-VID dataset. The performance gain is gradual and consistent by adding the temporal gradient first, the x-coordinate second, and by using the Lab channels last. The rank-1 accuracy increase reaches a maximum of 7.1% for iLIDS-VID and 3.3% for PRID2011 from baseline GOG to GOG3D. Since iLIDS-VID is the more challenging among the two datasets, the enhancement brought by these changes is more significant. Particularly, exploiting the temporal gradient proves the most useful by providing 4% improvement in rank-1 accuracy on iLIDS-VID. This feature adds some discriminative information that is hardly affected by the most challenging re-id attribute, illumination variation.

4.3.4 Comparison to GOG

To further validate the advantage brought by employing GOG3D over GOG for video-based re-id, we also compare the two algorithms using four other state-of-the-art metric learning methods:

Dataset	**iLIDS-VID**			**PRID2011**		
Rank	1	5	20	1	5	20
GOG + KISSME	49.5	73.4	88.1	81.0	94.2	99.2
GOG3D + KISSME	**55.1**	**78.9**	**92.7**	**83.3**	**94.9**	**99.2**
GOG + kNFST	63.9	85.9	95	87.8	97.4	99.9
GOG3D + kNFST	**74.3**	**92.2**	**98.9**	**89.6**	**97.8**	**100**
GOG + kLFDA	54.6	85.5	97.5	82.8	**96.8**	99.6
GOG3D + kLFDA	**67.4**	**90.7**	**98.8**	**85.1**	96.5	**99.9**
GOG + kMFA	56.3	85.3	97.6	83.3	**96.7**	99.5
GOG3D + kMFA	**66.5**	**90.9**	**98.9**	**85.4**	96.6	**99.9**

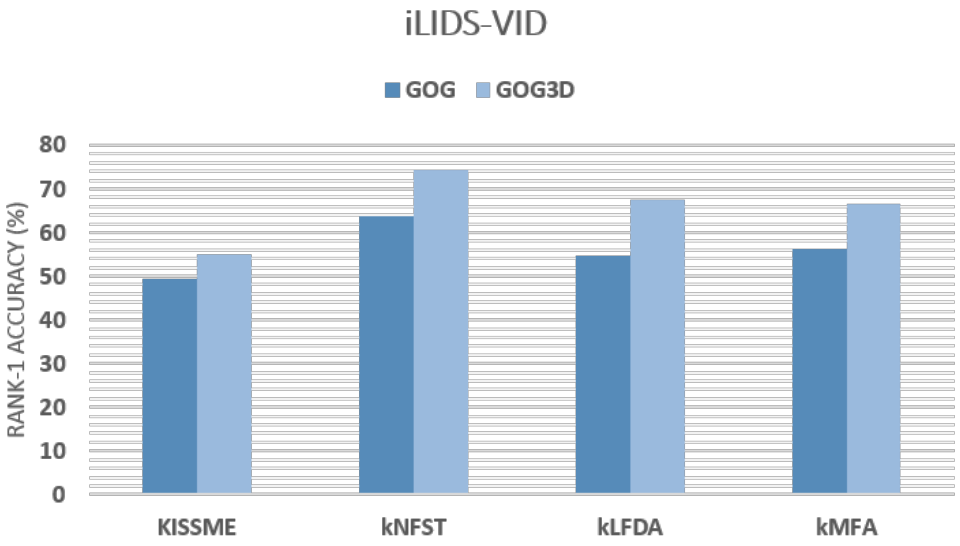
Table 4.2: Comparison to GOG. Best results in top-matching rates for each distance metric are in bold.

KISSME [85], kNFST [108], kLFDA and kMFA [54]. Here we also use the same parameter setting used in GOG while only changing the XQDA matching method. The results obtained can be seen in Table 4.2. For their ease of read, bar charts of rank-1 accuracy for both datasets are also shown in Figure 4.7.

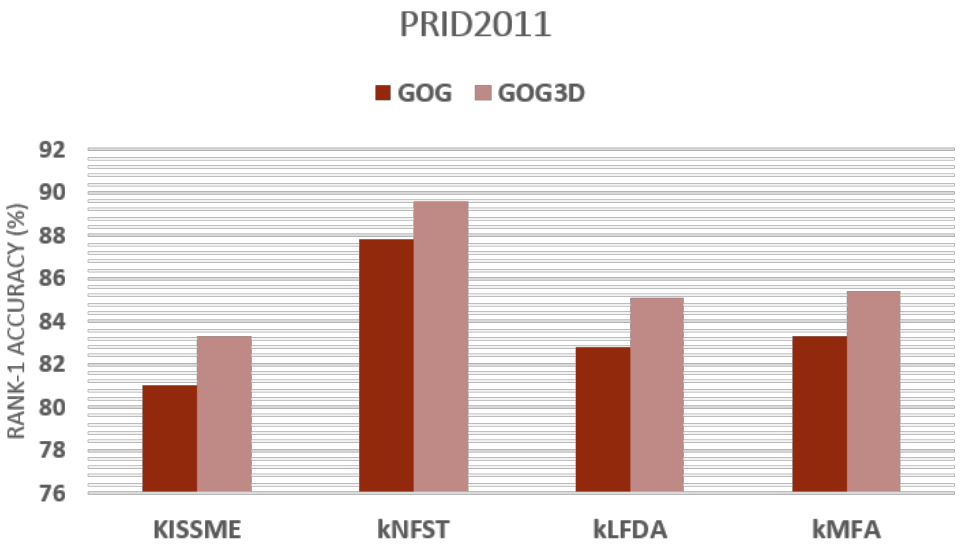
This evaluation further justifies the use of GOG3D over GOG for the video-based scenario as the improvement GOG3D brings especially in rank-1 accuracy is consistent for both datasets and for the four matching methods adopted. The improvement margin undergoes some fluctuations when employing different distance metrics. It reaches a maximum of almost 13% with kLFDA on iLIDS-VID and 2% on PRID2011 with most other distance metrics. The minimum rise in rank-1 accuracy on iLIDS-VID is 5.6% using KISSME method. Although the amount of accuracy increase depends on the characteristics of the dataset and the ability of the matching method to exploit the most convenient features in the learnt subspace, the consistent enhancement GOG3D brings makes it evidently more suited for video re-id.

4.3.5 Comparison to the State of the Art

The existing spatio-temporal descriptors such as ColHOG3D [33] and STFV3D [34] employ walking cycle extraction. This produces several fragments for each person sequence which requires special methods handling multi-shot matching to be jointly used. Consequently, employing them with the same evaluation protocol as ours where only one feature vector is obtained for each sequence is unfair and might cause their performance to downgrade. For this reason, video-based re-id systems are compared to our method in their original setting. To ensure a fair comparison,



(a)



(b)

Figure 4.7: Bar charts comparing GOG3D to GOG in rank-1 accuracy using various distance metrics.

Dataset	**iLIDS-VID**			**PRID2011**		
Rank	**1**	**5**	**20**	**1**	**5**	**20**
ColHOG3D+DVR [33]	39.5	61.1	81.0	40.0	71.7	92.2
STFV3D+KISSME [34]	44.3	71.7	91.7	64.1	87.3	92.0
RNN [141]	58	84	96	70	90	97
CNN+XQDA [148]	53.0	81.4	95.1	77.3	93.5	99.3
ASTPN [142]	62	86	98	77	95	99
DRAH [14]	64.0	86.0	96.3	88.7	97.9	99.7
PAM+KISSME [145]	**79.5**	95.1	99.1	92.5	**99.3**	**100**
SPW [138]	69.3	89.6	98.2	83.5	96.3	**100**
GOG3D^{10}+kNFST (proposed)	**80.0**	**95.3**	**99.5**	**93.6**	**99.4**	**100**
GOG3D^{15}+kNFST (proposed)	**79.5**	**95.4**	**99.5**	**94.0**	99.1	**100**

Table 4.3: Comparison to state-of-the-art methods. Best and second best results in top-matching rates are in bold.

the same evaluation protocol employed by most algorithms is adopted. More specifically, only sequences from 178 persons consisting of more than 27 frames are retained in the PRID2011 dataset. This protocol was initially employed by the methods requiring a minimum number of frames to extract walking cycles such as [33] and [34]. The standard evaluation protocol is also employed here where datasets are divided into half for training and half for testing, and the experiments are repeated over 10 trials. Average results in CMC top-matching rates are reported in Table 4.3.

The kNFST distance metric is used in this experiment. Unlike other distance metrics which attempt to minimise the within-scatter and maximise the between-scatter in the learnt subspace, kNFST collapses positive examples into a single point thus reducing the within-scatter into zero while ensuring a positive between-scatter. This is the reason why kNFST is more discriminative than other methods. As previously noted, an optimal setting is employed for GOG3D in terms of patch size, regularisation parameter and number of horizontal stripes (10 or 15). Hence, the results with kNFST are different from those reported in Table 4.2.

ColHOG3D [33] and STFV3D [34] are spatio-temporal descriptors that fall in the same category with GOG3D. However, GOG3D significantly outperforms them even when using the same distance metric KISSME as shown in the second row of Table 4.2. The gap in rank-1 accuracy with the better performing STFV3D is over 10% on iLIDS-VID and around 20% on PRID2011.

When compared to deep learning techniques, GOG3D+kNFST surpasses RNN [141], CNN+XQDA

[148] and ASTPN [142] by at least 18% on iLIDS-VID and 17% on PRID2011 in terms of rank-1 accuracy. It also outperforms multi-shot ranking methods DRAH [14] and SPW [138] by around 10% on iLIDS-VID for SPW and around 5% on PRID2011 for DRAH. The only method that exhibits comparable or slightly worse performance than GOG3D on both datasets is PAM+KISSME [145]. However, while not being in the same category with GOG3D, PAM requires (i) extracting low-level features, (ii) fitting GMMs to person sequences with many parameters to learn, and (iii) the final person representation does not fall in a Euclidean space. Hence, special methods should be devised for matching which limits the flexibility of its use with common off-the-shelf distance metrics.

In addition to impressive performance, GOG3D is simple, flexible and computationally efficient. The low computational cost is derived from the omission of additional tasks like walking cycle extraction and fragment selection used by similar space-time descriptors. It also avoids feature clustering and frame weighting required for multi-shot ranking methods that may involve further learning such as DRAH [14] and SPW [138]. Moreover, the high computational cost and hardware requirements needed to train deep neural networks are also spared. Finally, the flexibility of GOG3D feature is granted by the possibility of its use with any matching method in both supervised and unsupervised settings as it does not involve any learning.

4.3.6 Computational Cost

GOG3D is implemented in MATLAB and experiments are run on a desktop PC equipped with Intel Xeon X5550 @2.67GHz CPU. The average time to extract GOG3D features per frame is 0.44 seconds. It is computed on 10 video sequences from PRID2011 dataset and averaged over the number of frames constituting these sequences. Under the same setting, the time taken to compute GOG features is 0.35 seconds per frame. It is intuitive for GOG3D to be slightly slower than GOG since it uses additional pixel features and more horizontal stripes. However, compared to other video person re-id descriptors [33, 34], GOG3D is more efficient since it omits the walking cycle extraction step and any further post-processing. Moreover, frame-wise feature pooling employed with GOG3D renders the matching process very efficient. For instance, the average time taken to train the kNFST metric on PRID2011 dataset over 10 trials is 0.034 seconds, and the testing time on the same dataset is 0.008 seconds which is exceptionally fast for video-based re-id

methods.

4.4 Summary

A novel spatio-temporal descriptor for video-based person re-id that extends the 2-dimensional GOG feature into 3 dimensions was presented in this chapter. The proposed algorithm leverages the temporal correlation among the frames of a video sequence using the magnitude of the temporal gradient. It also benefits from other pixel features and regional settings that consistently boost the performance. The proposed GOG3D was also combined with the best existing supervised distance metric learning methods to produce a very impressive performance. Unlike other related space-time descriptors, by simply averaging frame-wise feature vectors over the person's sequence, robust representations can be achieved and can be easily fed into most off-the-shelf distance metrics.

A thorough analysis of the proposed descriptor was conducted on two widely used benchmarks using various distance metrics, thus highlighting the advantages brought by exploiting temporal cues. Extensive experiments showed that the performance achieved surpasses similar methods by a large margin. It also outperforms a number of existing deep learning and multi-shot ranking techniques. Since GOG3D is learning-free, it can be easily exploited in an unsupervised setting which is the topic of the next Chapter.

Chapter 5

Unsupervised Tracklet Matching for Video-based Re-Id

5.1 Introduction

GOG3D descriptor for video-based person re-id was proposed in the previous chapter. Despite being fully unsupervised and learning-free, it was employed in a supervised setting in conjunction with distance metric learning. The wide use of distance metric learning [11, 54, 85, 108] stems from its ability to efficiently boost the performance especially when the available annotated data is limited. When there is sufficient data, deep learning can be employed as an alternative [180, 182–184, 197]. Both of these methods have largely contributed into the advancement of the field by improving re-id performance. However, their applicability to real-life problems is still questionable. In a real-world setting, sufficient annotated data from concerned camera views is usually not available to train the model before re-id can take place. In addition to the costly and tedious annotation process, the availability of enough positive examples from the camera views in question is not always guaranteed, especially when they do not exhibit any adjacency or overlap but they are rather significantly farther apart. For these reasons, unsupervised re-id has drawn a lot of attention recently [135, 147, 155, 179, 180, 198] as a more realistic scenario. To move a step closer into solving the real-world problem, unsupervised video person re-id is investigated in this chapter.

One way to tackle multi-shot or video-based re-id is by adapting the task into the single-shot scenario. This is usually achieved by combining frame-wise feature vectors of a person tracklet into a single vector through average- or max-pooling [36, 53, 146–148]. In this case, an off-the-shelf matching method can be adopted to measure the similarity between probe and gallery vectors, as was done in Chapter 4. Despite its simplicity and efficiency, this approach suffers from some drawbacks. Firstly, the final representation of a person sequence is biased towards the pose that is most common among the frames constituting that sequence. Due to disjoint camera views causing significant illumination, pose and viewpoint angle variations among probe-gallery matches [35, 53, 161], this will frequently result in large intra-class variations which will degrade the performance. Secondly, such an approach treats all frames equally. Thus, noisy uninformative frames are assigned the same weight as the clear informative frames. This affects the quality and the discriminative power of the person's final representation as it becomes blighted by noise, or farther from its correct match in the feature space.

To address the disadvantages presented by feature pooling, we propose a simple yet robust distance measure based on the Naive Bayes Nearest Neighbour (NBNN) classifier [199]. The latter has been initially used in image classification in the context of Image-to-Class distance. However, by regarding person re-id as a classification problem, we show how the mathematical background of the NBNN algorithm can be leveraged to develop a set-based matching distance that is suited for the task. Moreover, we integrate the Spearman distance based on rank vectors into this framework as opposed to common distance metrics that are frequently used in unsupervised re-id such as the Euclidean or Cosine distance. The work related to unsupervised video re-id is first addressed in this chapter. Details of the proposed method are then presented, and the experiments carried out to validate the algorithm are finally described and analysed.

5.2 Related Work

The work related to supervised video-based person re-id was thoroughly discussed in Section 2.4. In this section, the existing algorithms on unsupervised video re-id are examined since they are tightly related to the method presented in this chapter.

In [34], the STFV3D spatio-temporal descriptor described in Section 2.4.1 was proposed. Al-

though computing the descriptor does not require any supervision *per se*, it was combined with supervised distance metric KISSME [85] for matching. Additionally, it was evaluated in an unsupervised manner using the nearest neighbour classifier. In the experimental work to follow in this chapter, the unsupervised version of STFV3D is used for benchmarking.

Ma *et al.* [200] on the other hand developed a video representation based on spatio-temporal pyramids using HOG3D features. Videos were sliced spatially and temporally, and the matching was performed using a modified dynamic time warping algorithm denoted Multi-Dimensional Time Shift Dynamic Time Warping (MDTS-DTW). MDTS-DTW jointly performs sequence alignment and chooses the optimal aligned segments in the final distance.

Khan *et al.* [196] proposed a person signature based on GMMs to describe pedestrians, but only used the mean information of the fitted GMMs. Subsequently, cross-view labels were automatically generated by studying the distribution of pairwise distances. These labels along with their signatures were then fed into the KISSME metric yielding the Unsupervised KISSME (Un-KISSME) method. In a subsequent algorithm denoted Part-Appearance Mixture (PAM) [145] (see Section 2.4.1), the signature was improved to incorporate both covariance and mean information from GMM components. For unsupervised matching, an f-Divergence based distance, Jeffrey's Divergence (JDiv), was used to measure the similarity between two GMMs.

Different from these methods, Liu *et al.* [155] and Ye *et al.* [179] devised techniques to estimate the labels progressively through iterative algorithms with hand-crafted features such as LOMO and GOG. In [155], a Stepwise Metric Promotion (SMP) approach was developed where ranking lists generated are refined through negative mining. In [179], graphs were constructed for samples in each camera and a Dynamic Graph Matching (DGM) model was proposed to estimate cross-camera labels. This is achieved by iteratively updating the estimated labels and the graphs. These methods implicitly [155] or explicitly [179] used labels from one camera view for model initialisation.

More recently, Chen *et al.* [147] and Li *et al.* [180] explored end-to-end deep learning architectures to associate within-camera and cross-camera tracklets by optimising specifically tailored objective functions. In [147], two margin-based association loss functions are jointly optimised constraining each frame to be associated with its optimal within-camera and cross-camera track-

lets. The algorithm is denoted as Deep Association Learning (DAL). A similar idea is used in Tracklet Association Unsupervised Deep Learning (TAUDL) method presented in [180] but with different loss functions. Moreover, within-camera tracklet labels are required for model initialisation. These are estimated using spatio-temporal information that is only available when raw videos are provided. It additionally depends on the viewing angle that dictates the time a subject can be perceived within the same camera view, and thus does not scale well to various camera views.

Nearest neighbour based methods have been successfully used in image classification [199, 201], action recognition [202, 203] and most recently they have been investigated in person re-id [155, 179, 204, 205]. For the re-id problem, these methods mainly exploit nearest neighbour information to optimise the ranking of the gallery list generated for a given probe by neighbourhood analysis of highly-ranked gallery elements [204, 205], or they employ neighbourhood information during the label estimation phase to progressively improve the output [155, 179]. Different from these works, label estimation and re-ranking are not performed in the proposed framework. Alternatively, nearest neighbour information is leveraged to design a set-based distance measure based on rank vectors.

5.3 Proposed Approach

Considering the large amount of video data currently available from surveillance cameras, existing pedestrian detection and tracking algorithms [206, 207] can be readily employed to extract person tracklets. For the supervised video re-id task, person tracklets are collected and each person is assigned an ID. Accordingly, within- and between-camera tracklets are annotated, and the tracklet's label is passed to its constituent frames. Each frame is then associated with a label representing the identity of the person involved. Some previous works on unsupervised video re-id make use of tracklet-wise labels available from one camera view for model initialisation [155, 179]. These methods are still widely considered unsupervised since they ignore cross-view labels. In this work, each person tracklet is treated separately for feature extraction and clustering as will be detailed in the remainder of this chapter. However, no assumption on the identity of the person presented in each tracklet is made. Therefore, no type of supervision is required.

The proposed method can be briefly described as follows. Firstly, high-dimensional GOG3D features are extracted from person tracklets. The dimensions of these features are then reduced using the PCA algorithm. The obtained frame-wise features of each tracklet are subsequently clustered using k-means algorithm, and each cluster is represented by its centroid. Finally, the proposed distance measure is used to compute probe-gallery distances. At the moment of testing, when a probe person tracklet is presented to the system, pairwise distances between probe and all gallery tracklets are computed and gallery elements are ranked according to their distance from the probe. The components of the proposed approach including person representation and matching method are detailed in the following. A representative diagram can be seen in Figure 5.1.

5.3.1 Features

For pedestrian sequence representation, GOG3D descriptor that was proposed, explained, and validated in Chapter 4 is used. We briefly recall that GOG3D is a 3-dimensional extension of GOG [12] that leverages temporal cues in addition to shape and colour information. In this work, the features are extracted separately from the frames of the sequences without performing any sort of pooling. After mean removal and ℓ_2-normalisation are applied to the extracted features, their dimension is reduced using the PCA algorithm. Reducing the dimension of the features is necessary in this case as no subspace will be learnt during the matching process. Hence, the curse of dimensionality problem will emerge due to highly correlated features which will cause the performance to drop.

Matching between frame-wise features directly is not only costly, it also ignores the rich representation obtained by combining information from various frames. For this purpose, PCA-reduced frame-wise features constituting each video tracklet are clustered using k-means algorithm [152]. Each cluster is finally represented by its centroid. This yields a set of feature vectors for each person tracklet that is equal to the number of clusters specified, thus rendering the probe-gallery matching a set-based process.

5.3.2 Proposed Distance Measure

The proposed unsupervised distance measure aims at computing the dissimilarity between two sets of feature vectors, the first set belonging to a probe person tracklet and the second to a gallery

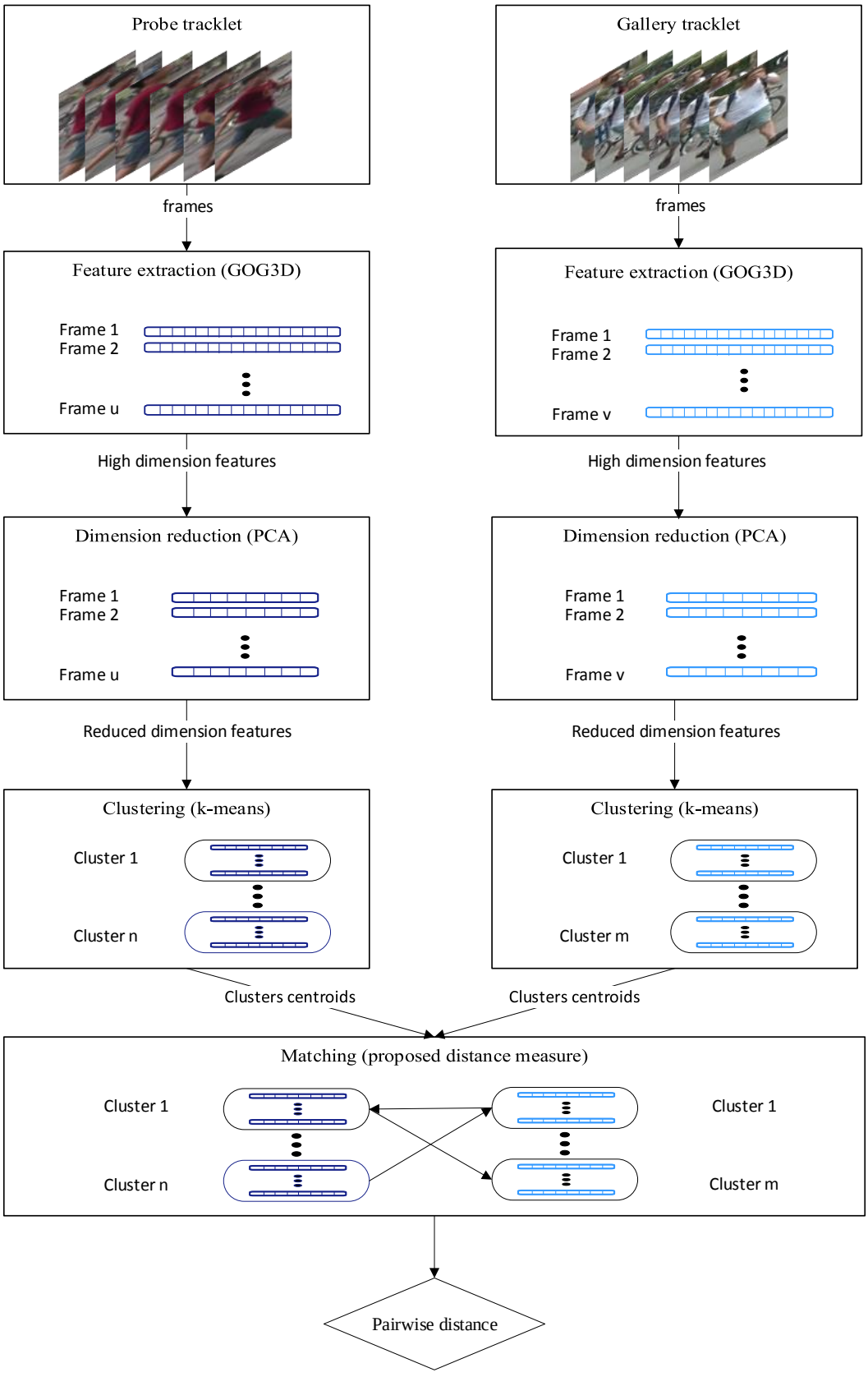


Figure 5.1: Diagram of the proposed unsupervised approach.

person tracklet. The mathematical formulation of this distance is derived as follows. Let $P = (p_1, \ldots, p_n)$ be a probe person sequence represented by n feature vectors obtained by clustering frame-wise features. These n vectors are in fact the clusters centroids output by the k-means algorithm. The problem boils down to finding the class in the gallery set to which this probe belongs. Namely, the purpose is to find the gallery subject $G_k = (g_1^k, \ldots, g_m^k)$ that is a correct match for query person P. The Maximum A Posteriori (MAP) classifier minimises the average classification error:

$$\hat{C} = \underset{k}{\operatorname{argmax}}\, p(G_k|P), \tag{5.1}$$

where $k \in \{1, \ldots, C\}$, C being the number of classes in the gallery set. The following Bayes' rule can subsequently be applied:

$$p(G_k|P) = \frac{p(G_k) \cdot p(P|G_k)}{p(P)}. \tag{5.2}$$

Assuming a uniform prior over classes $G_k, k \in 1, \ldots, C$, and $p(P)$ being a constant independent of the class G_k, MAP classifier reduces into the Maximum Likelihood (ML) classifier as such:

$$\hat{C} = \underset{k}{\operatorname{argmax}}\, p(G_k|P) = \underset{k}{\operatorname{argmax}}\, p(P|G_k). \tag{5.3}$$

Assuming that probe descriptors $p_1, \ldots, p_n$ satisfy the Naive Bayes assumption (they are independent and identically distributed (i.i.d.) given class G_k), $p(P|G_k)$ can be written as:

$$p(P|G_k) = p(p_1, \ldots, p_n|G_k) = \prod_{i=1}^{n} p(p_i|G_k). \tag{5.4}$$

Substituting $p(P|G_k)$ from Equation (5.4) in Equation (5.3) and taking the $\log$ probability yields:

$$\hat{C} = \underset{k}{\operatorname{argmax}} \log p(P|G_k) = \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(p_i|G_k). \tag{5.5}$$

If $g_1^k, \ldots, g_m^k$ are all the descriptors in class G_k, then $p(p_i|G_k)$ can be approximated using a Parzen window estimator [201] with similarity kernel K by:

$$\hat{p}(p_i|G_k) = \frac{1}{m} \sum_{j=1}^{m} K(p_i - g_j^k), \tag{5.6}$$

$\hat{p}(p_i|G_k)$ in Equation (5.6) can be further approximated by taking the r largest elements in this summation. They correspond to the r nearest neighbours of p_i in class G_k:

$$\hat{p_r}(p_i|G_k) = \frac{1}{m}\sum_{j=1}^{r} K(p_i - g_j^k). \tag{5.7}$$

This can be taken to the extreme by using a single nearest neighbour of p_i in $G_k = \{g_1^k, \ldots, g_m^k\}$ denoted $NN_{G_k}(p_i)$. Hence,

$$\hat{p_1}(p_i|G_k) = \frac{1}{m}K(p_i - NN_{G_k}(p_i)). \tag{5.8}$$

Choosing a single nearest neighbour is particularly appealing because in this case, Equation (5.5) can reduce into a very simple format [201]. For instance, by selecting a Gaussian kernel for K and combining Equations (5.5) and (5.8), we obtain:

$$\begin{aligned}
\hat{C} &= \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(p_i|G_k) \\
&= \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log(\frac{1}{m}K(p_i - NN_{G_k}(p_i)) \\
&= \underset{k}{\operatorname{argmax}} \sum_{i=1}^{n} \log(\frac{1}{m}e^{-\frac{1}{2\sigma^2}||p_i - NN_{G_k}(p_i)||^2}) \\
&= \underset{k}{\operatorname{argmin}} \sum_{i=1}^{n} ||p_i - NN_{G_k}(p_i)||^2.
\end{aligned} \tag{5.9}$$

In other terms, this classification rule entails finding the gallery instance with the minimum distance from the probe. Based on the previous analysis, noting that in our case $n = m = 5$ (number of k-means clusters) for all probe and gallery elements, the distance between a probe instance $P = (p_1, \ldots, p_n)$ and a gallery instance $G = (g_1, \ldots, g_m)$ is defined as:

$$d_{P \to G} = \sum_{i=1}^{n} \delta(p_i, NN_G(p_i)), \tag{5.10}$$

where $NN_G(p_i)$ is the nearest neighbour of p_i in G, and δ is a distance measure such as the Euclidean distance. Similarly, the distance from G to P can be computed as:

$$d_{G \to P} = \sum_{i=1}^{m} \delta(g_i, NN_P(g_i)), \tag{5.11}$$

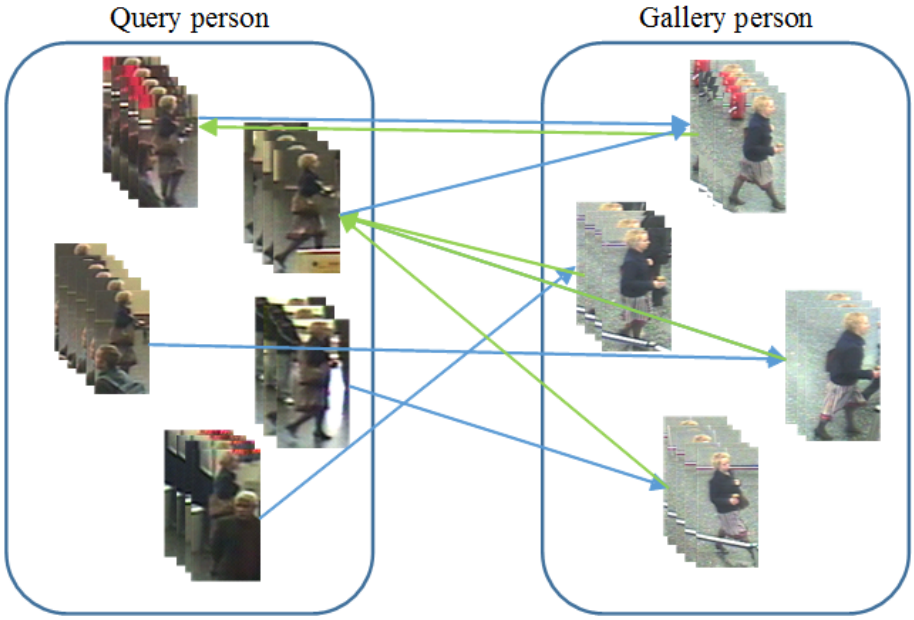


Figure 5.2: A set-based matching process allows the exclusion of outliers and the selection of better representative frames during matching. The arrows lead a cluster to its nearest neighbour.

and the final similarity score between P and G combines both formulas in a symmetric manner,

$$d_{\mathrm{NN}}(P,G) = d_{P\to G} + d_{G\to P}. \tag{5.12}$$

Based on this definition, the correct match for a probe P consists of the gallery element G with the minimum distance from P. It is worth noting here that a more general derivation incorporating more than one nearest neighbour for each descriptor p_i with arbitrary values for n and m is given by:

$$d_{P\to G} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{r}\delta(p_i, g_j), \tag{5.13}$$

where $g_j, j = 1,\ldots,r$ are the r nearest neighbours of p_i in class G. Similarly,

$$d_{G\to P} = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{r}\delta(g_i, p_j), \tag{5.14}$$

where $p_j, j = 1,\ldots,r$ are the r nearest neighbours of g_i in probe P. Eventually, the final distance d_{NN} is derived as in Equation (5.12). A diagram depicting the proposed distance measure can be seen in Figure 5.2.

5.3.3 Embedded Matching Distance

The embedded distance δ plays a crucial role in the success of the proposed re-id system. Although the Euclidean distance has been widely used with unsupervised person re-id methods [10, 147], it still has its shortcomings. In addition to losing its discriminability with high-dimensional data as discussed in Section 2.3.2, it is very vulnerable to outliers since it treats all features with equal importance. For instance, notable fluctuations in a few features might affect the final pairwise distance drastically which will cause misidentification issues as the intra-class distance becomes bigger than some inter-class distances. Since these fluctuations are likely to happen in the re-id scenario due to cross-view camera variations and repeated occlusions, we propose to use a rank-based distance measure instead. For this purpose, the Spearman distance is used. It is equivalent to the Pearson correlation distance when the latter is applied to rank vectors. In other words, feature vectors $X = (x_1, \ldots, x_d)$ and $Y = (y_1, \ldots, y_d)$ are converted into rank vectors $r_X = (r_1^X, \ldots, r_d^X)$ and $r_Y = (r_1^Y, \ldots, r_d^Y)$ by replacing all the values by their respective ranks. For instance, if vector $X = (0.7, 0.2, 0.4)$, then $r_X = (3, 1, 2)$. Subsequently, the Spearman distance between X and Y is computed in terms of Spearman rank correlation coefficient ρ_s as follows:

$$d_{\mathrm{S}}(X,Y) = 1 - \rho_s(X,Y) = 1 - \frac{\sum_{i=1}^{d}(r_i^X - \overline{r_X})(r_i^Y - \overline{r_Y})}{\sqrt{\sum_{i=1}^{d}(r_i^X - \overline{r_X})^2}\sqrt{\sum_{i=1}^{d}(r_i^Y - \overline{r_Y})^2}}, \tag{5.15}$$

where $\overline{r_X} = \frac{1}{d}\sum_{i=1}^{d} r_i^X = \frac{d+1}{2}$ and $\overline{r_Y} = \frac{1}{d}\sum_{i=1}^{d} r_i^Y = \frac{d+1}{2}$ are the means of rank vectors r_X and r_Y, respectively.

To highlight the advantages brought by the use of the Spearman distance, we define other common distance measures: the Euclidean, Cosine, and Pearson correlation distance. Element-wise notation is used for more clarity. If $\overline{x}$ and $\overline{y}$ are the respective means of vectors X and Y, then the distances in the previous order are defined as follows:

$$d_{\mathrm{Euc}}(X,Y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}, \tag{5.16}$$

$$d_{\mathrm{Cos}}(X,Y) = 1 - \frac{\sum_{i=1}^{d} x_i y_i}{\sqrt{\sum_{i=1}^{d} x_i^2}\sqrt{\sum_{i=1}^{d} y_i^2}}, \tag{5.17}$$

$$d_{\mathrm{P}}(X,Y) = 1 - \frac{\sum_{i=1}^{d}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{d}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{d}(y_i - \overline{y})^2}}. \tag{5.18}$$

Since the features are mean-centered ($\overline{x} = \overline{y} = 0$), it can be easily seen that Cosine and Pearson distances become equivalent. Moreover, given that they are also ℓ_2-normalised ($\sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{\sum_{i=1}^{d} y_i^2} = 1$), it can be easily shown that the Euclidean and Cosine distance measures become equivalent. In fact,

$$d^2_{\text{Euc}}(X,Y) = \sum_{i=1}^{d} (x_i - y_i)^2 = \sum_{i=1}^{d} (x_i^2 - 2x_i y_i + y_i^2) = 2 - 2\sum_{i=1}^{d} x_i y_i = 2d_{\text{Cos}}(X,Y). \tag{5.19}$$

Therefore, the squared Euclidean distance is proportional to the Cosine distance. Since we are concerned about the relative order of pairwise distances rather than their exact values, both Euclidean and Cosine distances will generate the same ranking list of gallery elements given a certain probe, thus yielding the same results.

Based on the previous analysis, the comparison of the Spearman distance to one of these distance measures applies to all. In fact, from the formulations we can see that the Pearson distance applied to rank vectors equates to the Spearman distance. By taking ranks instead of raw values, the Spearman distance relaxes the assumption held by the former of a linear relationship between feature vectors for higher similarity, and looks for patterns of monotonicity instead. In other words, in this case two vectors exhibit higher similarity when their features undergo the same type of fluctuations without them being necessarily linearly correlated. This proves to be more useful in the case of person re-id where many challenges can arise causing outlying feature values to largely affect computed pairwise distances. The superiority of the Spearman distance compared to other metrics will be shown experimentally in section 5.4.3.

5.3.4 Implementation Details

For the feature extraction stage, the optimal setting of GOG3D described in Chapter 4 is employed. Particularly, the patch size is set to 9×9 pixels with 2 pixels extraction intervals, and the images are divided into 10 horizontal regions with 50% overlap. To ensure non-singular covariance matrices, the regularisation parameter is set to $\epsilon_0 = 0.0001$. The evaluation is conducted on three datasets: PRID2011, iLIDS-VID and MARS. PRID2011 and iLIDS-VID frames are kept in their original size of 128×64 pixels, and MARS frames are resized to 128×48 pixels before extracting the features for efficiency reasons. This results in frame-wise feature vectors of 22,780 dimensions

each. The number of k-means clusters is set to 5 clusters ($m = n = 5$ in Equations (5.13) and (5.14)) for each probe or gallery tracklet in all the experiments. In fact, 10 clusters were employed in [14], but since a small difference was empirically observed with a smaller k, $k = 5$ is conveniently used for higher computational efficiency. Similarly to [145], the dimension of the features is reduced using PCA so that enough components are kept to retain 95% of the variance in the data. One nearest neighbour is used in the following experiments ($r = 1$ in Equations (5.13) and (5.14)) unless otherwise specified.

5.4 Experiments and Results

5.4.1 Datasets and Evaluation Protocol

To evaluate the proposed unsupervised method, the same datasets used in the previous chapter are employed here, that is, PRID2011 [51] and iLIDS-VID [33] (see Section 4.3.1). Despite these datasets being very challenging, they are small in size and manually annotated which might not reflect the generalisation ability of our algorithm. Therefore, we additionally use the large-scale MARS dataset [148]. The evaluation protocol employed is the same adopted by other similar methods [33, 147, 155, 179, 180, 196] and the work presented in Chapter 4. Namely, for PRID2011 dataset, 178 video pairs including more than 27 frames per sequence are kept and the rest are discarded. Although training is not needed for this algorithm, like other similar methods, both PRID2011 and iLIDS-VID datasets are randomly divided into half for training and half for testing. This yields 89 persons in each subset for PRID2011 and 150 for iLIDS-VID. In fact, the training set is only used to compute the subspace learned by PCA to which the test set is subsequently projected. Experiments are repeated over 10 trials and average results in CMC top-matching rates are reported.

MARS [148] is a large-scale video benchmark for person re-id collected on a university campus by 6 near-synchronised cameras. It consists of 1,261 pedestrians each appearing in 2 cameras at least. Unlike other benchmarks, persons are not manually detected and cropped. Alternatively, a more realistic application is offered by automatic pedestrian detection and tracking using Deformable Part Models (DPM) [206] detector and Generalised Maximum Multi Clique Problem (GMMCP) [207] tracker. This results in a total of 20,478 person tracklets with an average of 13.2 tracklets

Table 5.1: Comparison against state-of-the-art results on PRID2011 and iLIDS-VID datasets in top-matching rates.

	Dataset	**iLIDS-VID**			**PRID2011**		
	Rank R	**R = 1**	**R = 5**	**R = 20**	**R = 1**	**R = 5**	**R = 20**
Unsupervised	**STFV3D [34]**	37.0	64.3	86.9	42.1	71.9	91.6
	MDTS-DTW [200]	31.5	62.1	82.4	41.7	67.1	90.1
	unKISS [196]	38.2	65.7	84.1	59.2	81.7	96.1
	PAM+LOMO [145]	33.3	57.8	80.5	70.6	90.2	97.1
	DGM+IDE [179]	36.2	62.8	82.7	56.4	81.3	96.4
	DGM+XQDA [179]	31.3	55.3	83.4	82.4	95.4	99.8
	SMP [155]	41.7	66.3	80.7	80.9	95.6	99.4
	DAL (ResNet50) [147]	56.9	80.6	91.9	85.3	**97.0**	**99.6**
	TAUDL [180]	26.7	51.3	82.0	49.4	78.7	98.9
	Proposed	**79.1**	**93.5**	**97.5**	**91.7**	96.7	98.7
Supervised	**Snippet [31]**	85.4	96.7	99.5	93.0	99.3	100.0
	QAN [140]	68.0	86.8	97.4	90.3	98.2	100.0
	STAN [139]	80.2	-	-	93.2	-	-
	SDM [144]	60.2	84.7	95.2	85.2	97.1	99.6

per person, including 3,248 distractors caused by false detection or tracking. Crops are resized to 128×256 pixels. In addition to background clutter and inconsistent lighting, the main challenges associated with MARS are the misalignment and change in scale at which persons are seen. This is caused by automatic detection and tracking. The standard train/test split [148] used by other algorithms [147, 155, 179, 180] is adopted for evaluation. It includes all the tracklets of 625 pedestrians in the training set and those of 636 pedestrians in the test set. One tracklet is selected for each person from each view as probe. As multiple ground truths correspond to each query, in addition to the CMC curve, mAP is used to evaluate the performance. Example images from MARS dataset can be seen in Figure 5.3.

5.4.2 Comparison to State of the Art

We compare our method against 9 state-of-the-art unsupervised person re-id methods on PRID2011 and iLIDS-VID datasets, and 5 on MARS dataset. We also report some recent supervised re-id results to show the current existing gap in performance. STFV3D [34], MDTS-DTW [200], UnKISSME [196] and PAM+LOMO [145] rely on hand-crafted representations with different unsupervised matching methods. SMP [155] and DGM [179] also leverage hand-crafted features, but design algorithms to generate labels that are subsequently used with supervised metric learning. DAL [147] and TAUDL [180] are unsupervised deep models that attempt to associate tracklets in

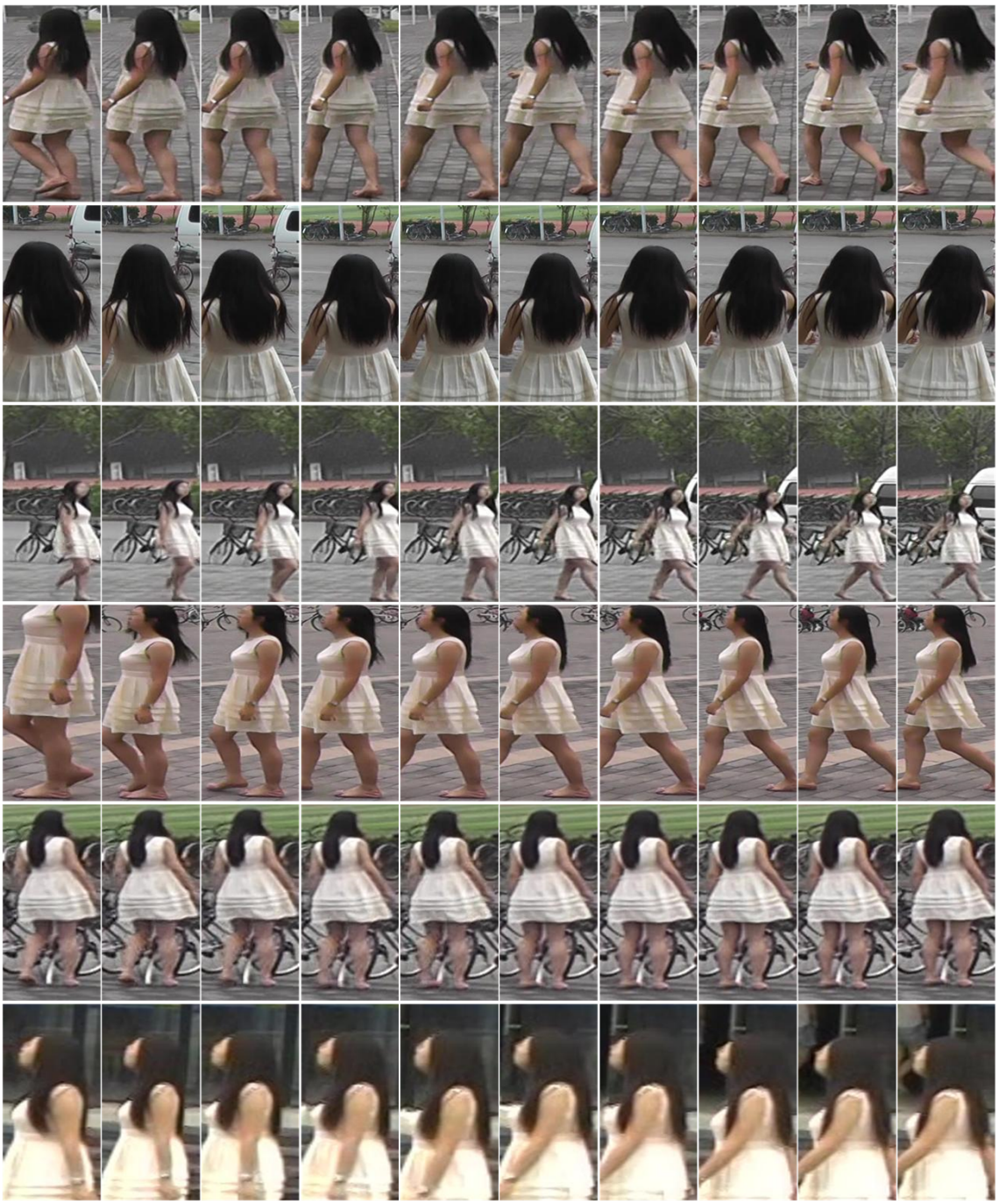


Figure 5.3: Example images from MARS dataset. The same subject is shown in all images. Each row includes consecutive frames from a person tracklet taken from a different camera views.

Figure 5.4: Significant misalignment between consecutive frames can be observed in some MARS tracklets which adds to the challenges.

an end-to-end fashion. It is clear from the results reported in Table 5.1 that the proposed method outperforms all existing unsupervised techniques on PRID2011 and iLIDS-VID datasets by a margin of approximately 6% and 22% respectively in rank-1 accuracy with its closest competitor DAL. It also surpasses supervised methods Sequential Decision Making (SDM) [144] and Quality Aware Network (QAN) [140], while the gap with the best performing supervised method [31] is almost 6% on iLIDS-VID and less than 2% on PRID2011. The proposed approach also achieves very competitive results on MARS benchmark as can be seen in Table 5.2. On the latter, it outperforms the 3 non-deep models in rank-1 accuracy and achieves competitive performance with the deep models DAL and TAUDL. However, in general the gap in performance between supervised and unsupervised methods is still very large (approximately 40%) on this dataset.

The poor quality of the bounding boxes produced by automatic detection and tracking on MARS which causes serious misalignment between consecutive frames, as can be seen in Figure 5.4, is problematic for hand-crafted features using rigid part-based models (horizontal strips). Furthermore, it is no surprise that deep learning methods can scale better to large-scale datasets while suffering with small ones as validated by the poor performance of TAUDL on iLIDS-VID and PRID2011. Nonetheless, the amount of data available at the moment of re-identification might not always be substantial, therefore a successful re-id system should be able to strike some balance between both scenarios.

Finally, the proposed approach contributes massively into closing the gap in rank-1 accuracy between supervised and unsupervised re-id on the small but challenging datasets PRID2011 and iLIDS-VID while competing with deep learning methods on the large-scale MARS dataset.

Table 5.2: Comparison against state-of-the-art on MARS dataset in top-matching rates and mean Average Precision.

	Dataset	**MARS**			
	Rank R	**R = 1**	**R = 5**	**R = 20**	**mAP**
Unsupervised	**DGM+IDE [179]**	36.8	54.0	68.5	21.3
	DGM+XQDA [179]	23.6	38.2	54.7	11.2
	SMP [155]	23.6	35.8	44.9	10.5
	DAL (ResNet50) [147]	**46.8**	**63.9**	**77.5**	21.4
	TAUDL [180]	43.8	59.9	72.8	**29.1**
	Proposed	39.7	53.2	64.1	20.1
Supervised	**Snippet [31]**	86.3	94.7	98.2	76.1
	STAN [139]	82.3	-	-	65.8
	PABR [156]	85.1	94.2	97.4	83.9
	SDM [144]	71.2	85.7	94.3	-

5.4.3 Algorithm Analysis

In this section, individual components of the proposed framework are analysed highlighting the improvement they bring upon the overall performance. For computational reasons and due to the lack of resources, this analysis is conducted solely on PRID2011 and iLIDS-VID.

Features

Although the superiority of GOG3D over GOG was extensively shown in Chapter 4 when combined with supervised distance metric learning, we further evaluate the advantages it brings in the proposed unsupervised setting. For this purpose, experiments were conducted involving GOG and GOG3D features on PRID2011 and iLIDS-VID datasets while keeping other components of the system intact. The results obtained are shown in Figure 5.5. Leveraging temporal information added to the other changes applied to the pixel feature vector and spatial image decomposition contribute into improving the performance by a margin of approximately 3% for PRID2011 and 14% for iLIDS-VID in rank-1 accuracy. The improvement on iLIDS-VID is more notable since this dataset additionally suffers from occlusions and significant illumination changes. Hence, leveraging motion information brings additional discriminative information into pedestrian representation.

Dimension Reduction

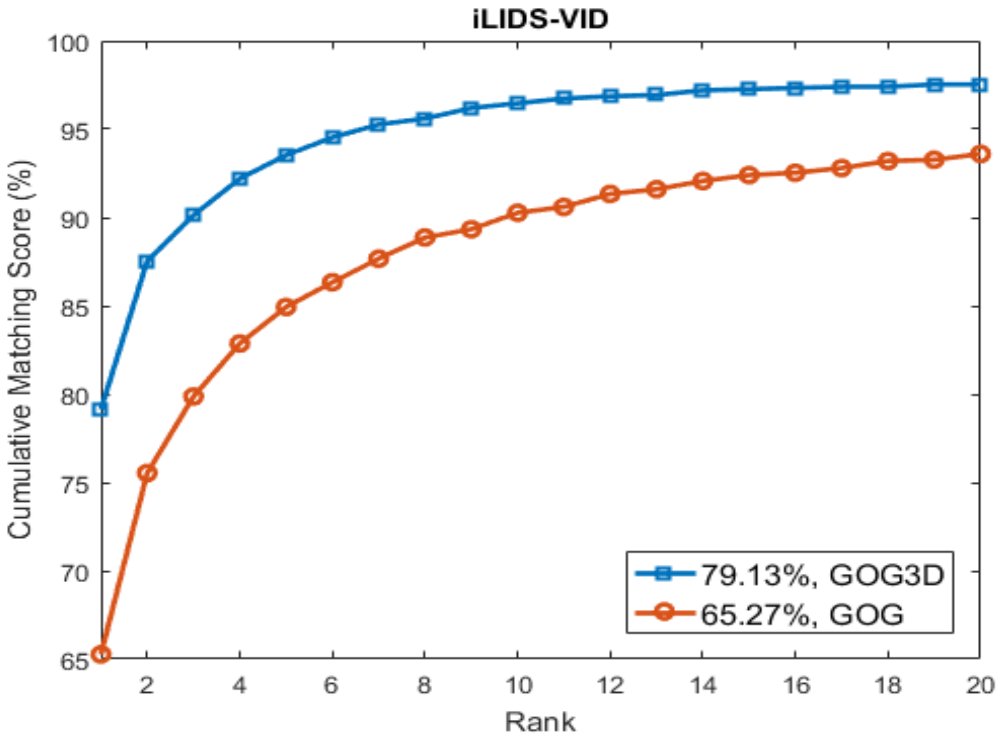
In addition to efficiency considerations, reducing the dimension of the feature vectors is essential when no supervision is involved to discard redundant features and select the most discriminative ones for re-identification. This step is particularly important in this case because rank vectors are used for matching. Ranking a high-dimensional feature vector where values slightly differ from each other is sub-optimal. Therefore, to gauge the effect this might have on the system's performance, the proposed algorithm is evaluated with and without PCA. The results obtained are shown in Figure 5.6. As expected, employing PCA before matching brings substantial improvement in rank-1 accuracy especially for iLIDS-VID dataset (around 40%), while the improvement for PRID2011 is around 15%. Given the challenges associated with this dataset, selecting discriminative features is essential to match rank feature vectors of the same individual.

Embedded distance

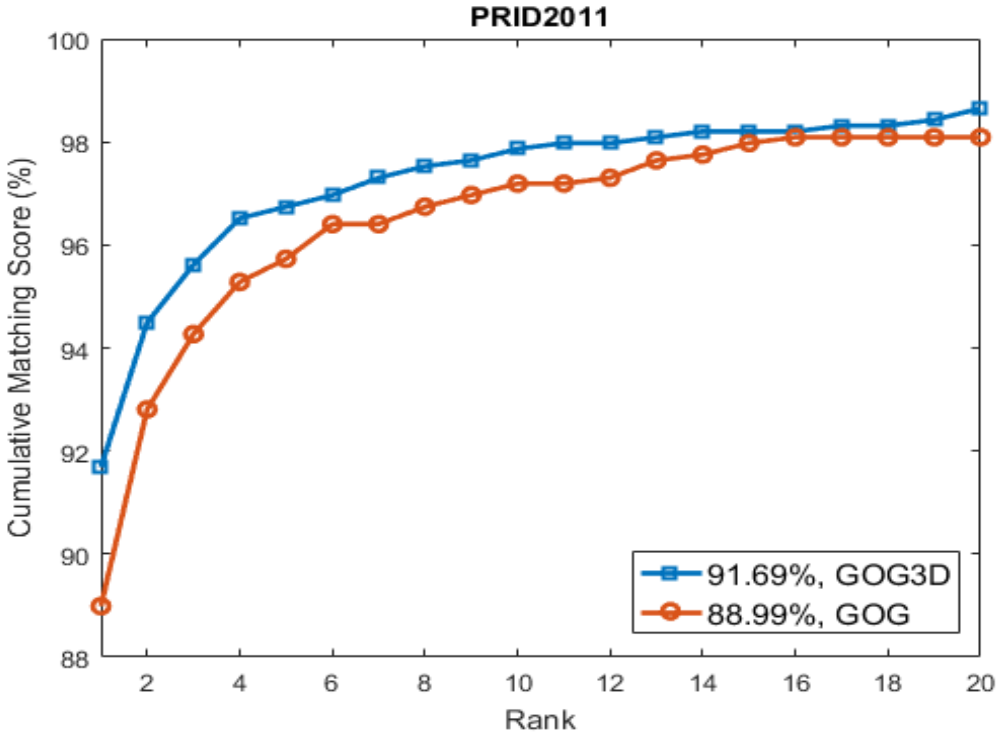
To evaluate the effect of the embedded distance on the re-id performance, results using two distance metrics, Cityblock (ℓ_1-distance) and Euclidean are also reported. The Cityblock distance between two vectors $X = (x_1, \ldots, x_d)$ and $Y = (y_1, \ldots, y_d)$ is defined as $\sum_{i=1}^{d} |x_i - y_i|$. As previously mentioned, since the extracted features are mean-centered and ℓ_2-normalised, the Cosine distance and the Pearson correlation distance are similar to the Euclidean distance (this was indeed verified experimentally). As can be seen in Figure 5.7, although the Euclidean distance consistently outperforms the Cityblock distance for both datasets by approximately 15% in rank-1 accuracy, the improvement with the Spearman distance is even more substantial. It reaches a maximum of approximately 47% in rank-1 accuracy compared to the Euclidean distance on iLIDS-VID dataset. Looking for monotonicity instead of linearity proved in fact useful by considerably improving the system's accuracy.

Set-based Matching

We also compare our method against a commonly used matching protocol for video person re-id, average-pooling [36, 53, 146, 180]. In this case, frame-wise feature vectors of a tracklet are averaged to obtain one representation of each person sequence. The single-shot scenario with the Spearman distance is subsequently used for matching. The results obtained are shown in

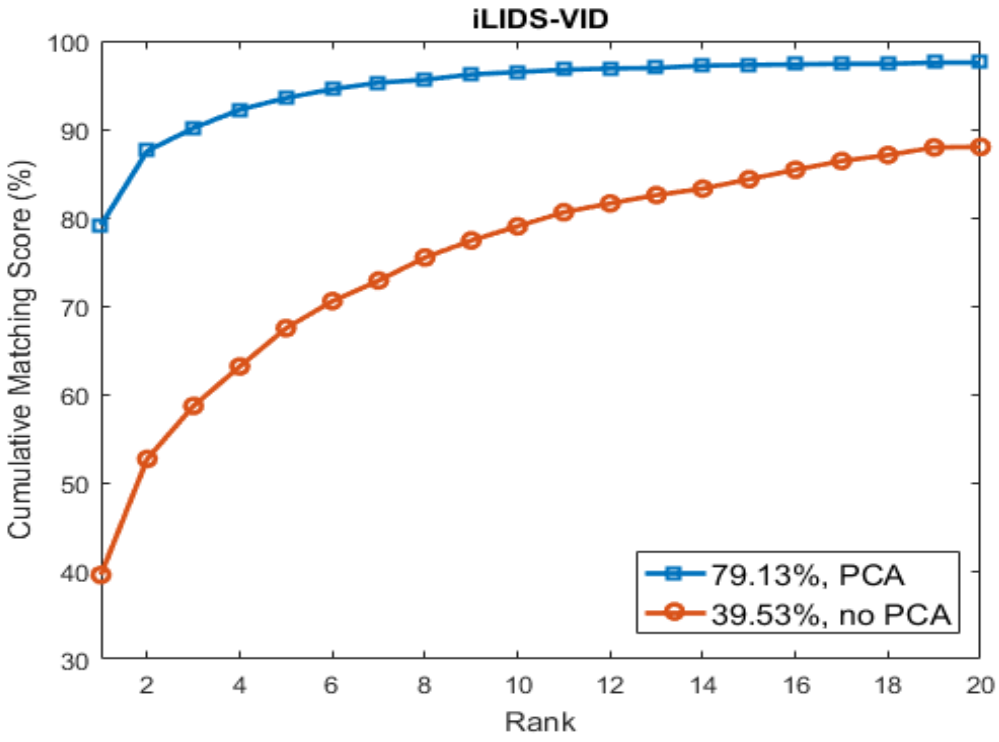


(a)

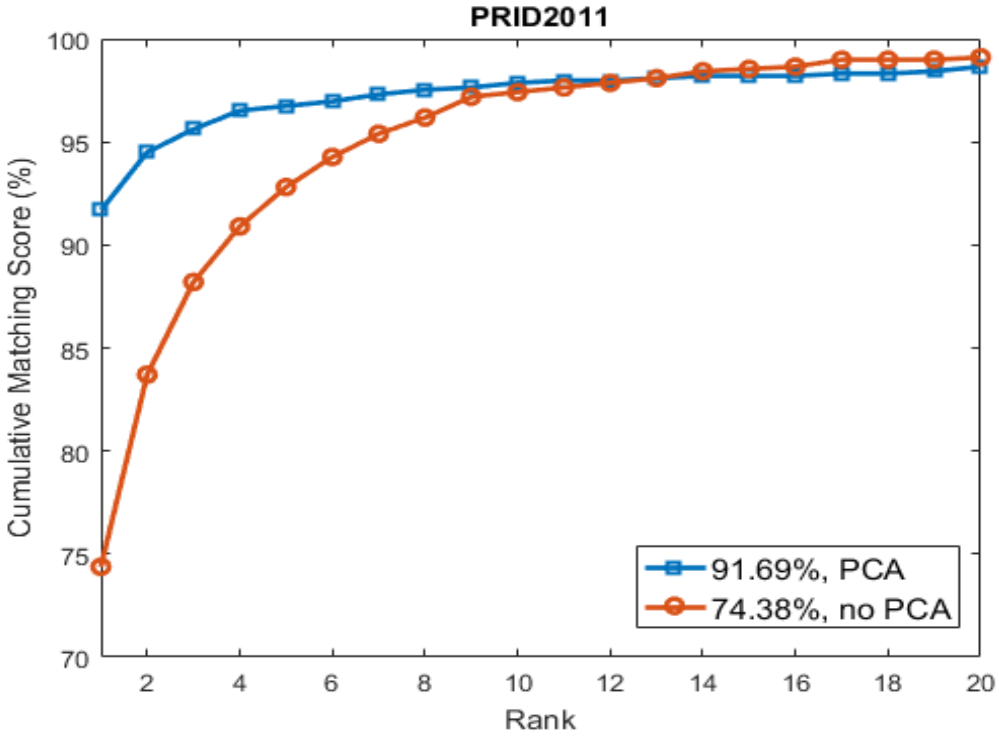


(b)

Figure 5.5: CMC curves using GOG and GOG3D features on iLIDS-VID and PRID2011 datasets.

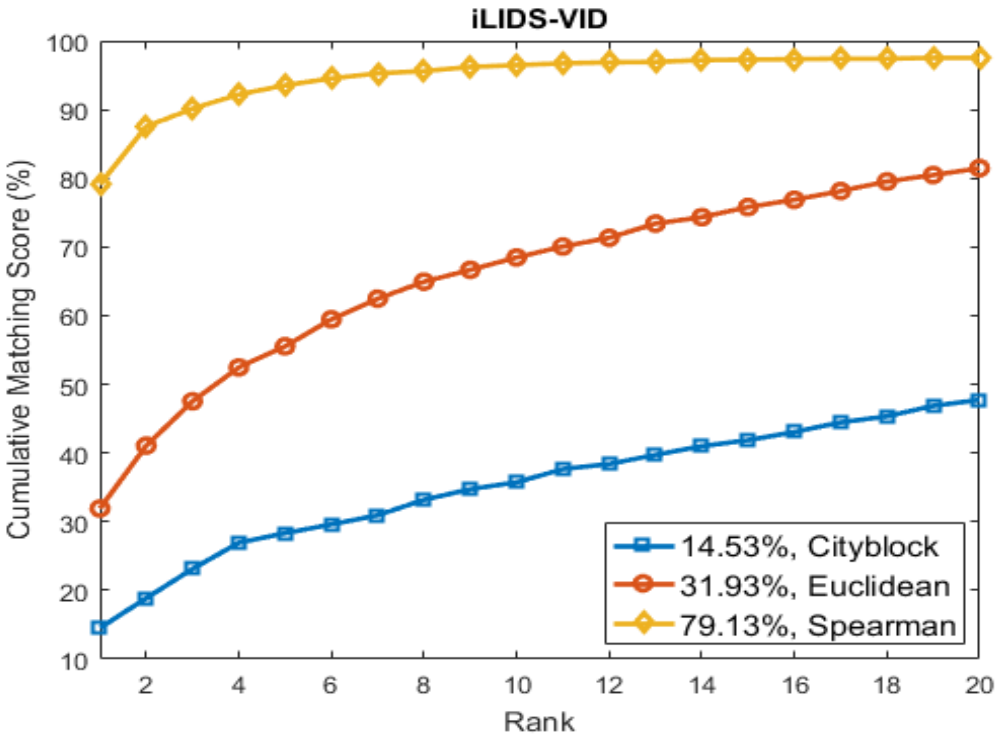


(a)

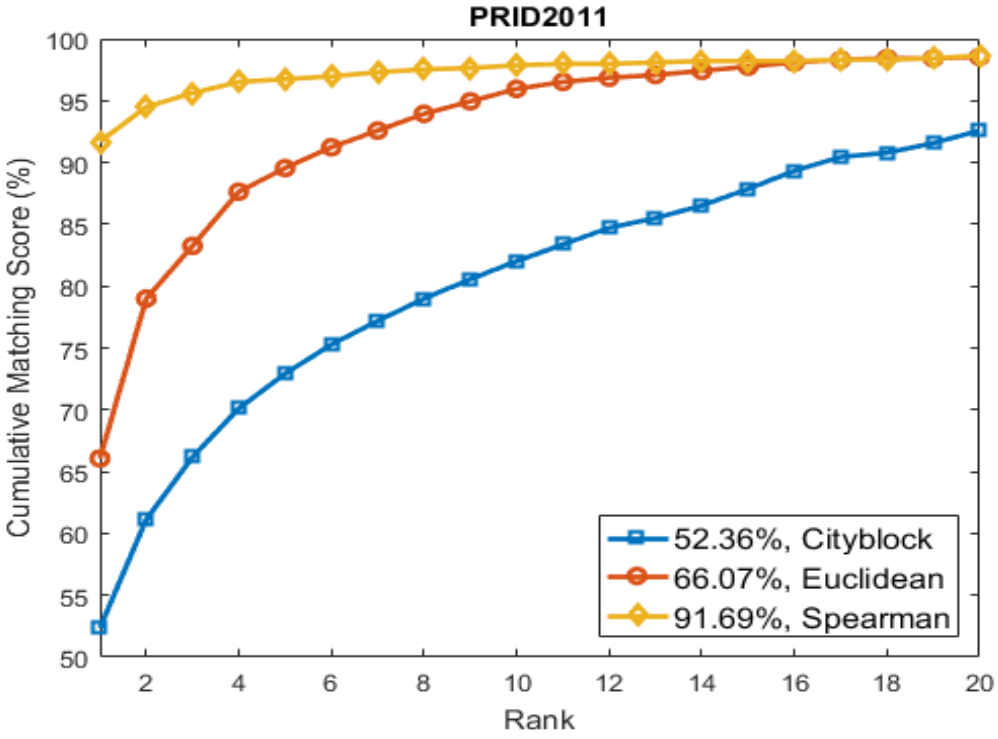


(b)

Figure 5.6: CMC curves of the results with and without performing PCA on iLIDS-VID and PRID2011 datasets.

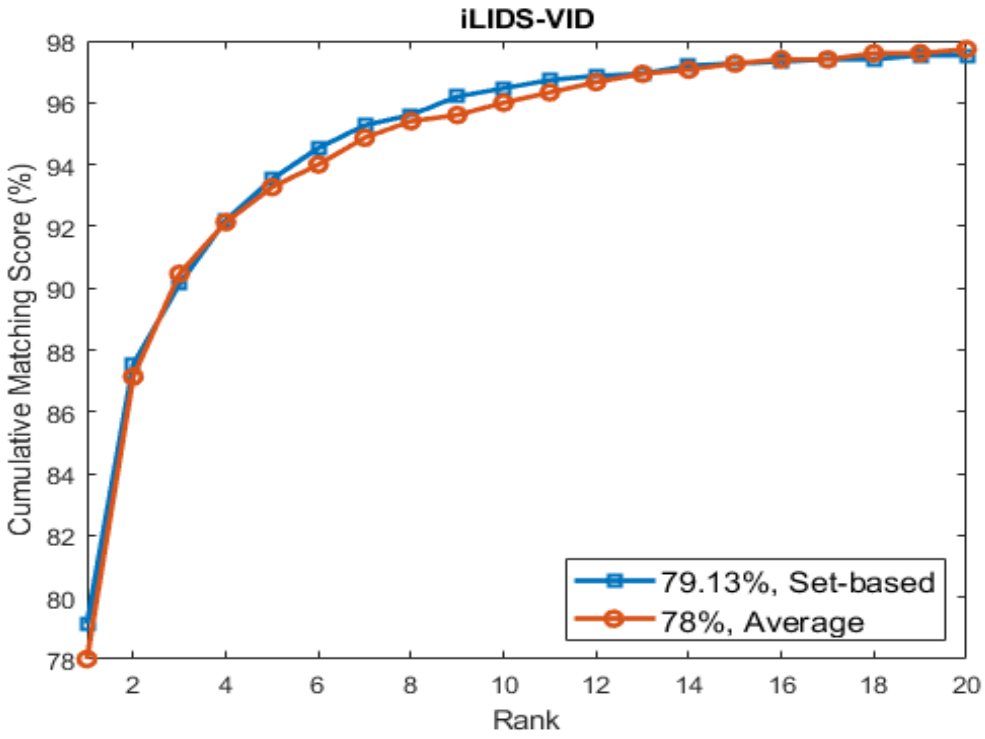


(a)

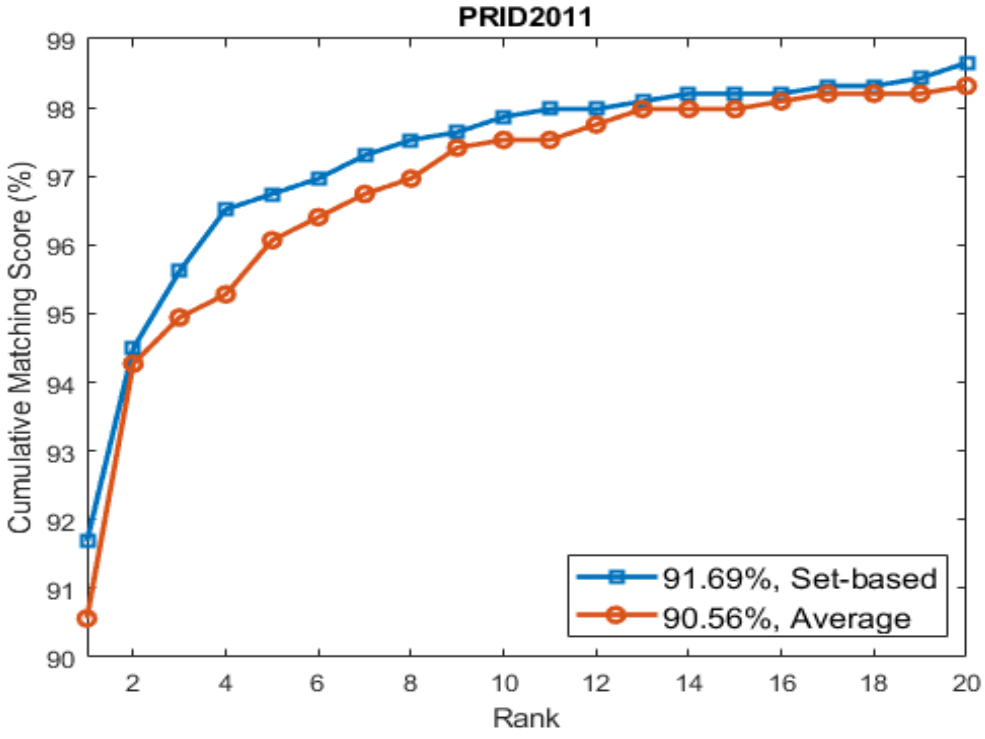


(b)

Figure 5.7: CMC curves using Cityblock, Euclidean and Spearman distance measures on iLIDS-VID and PRID2011 datasets.

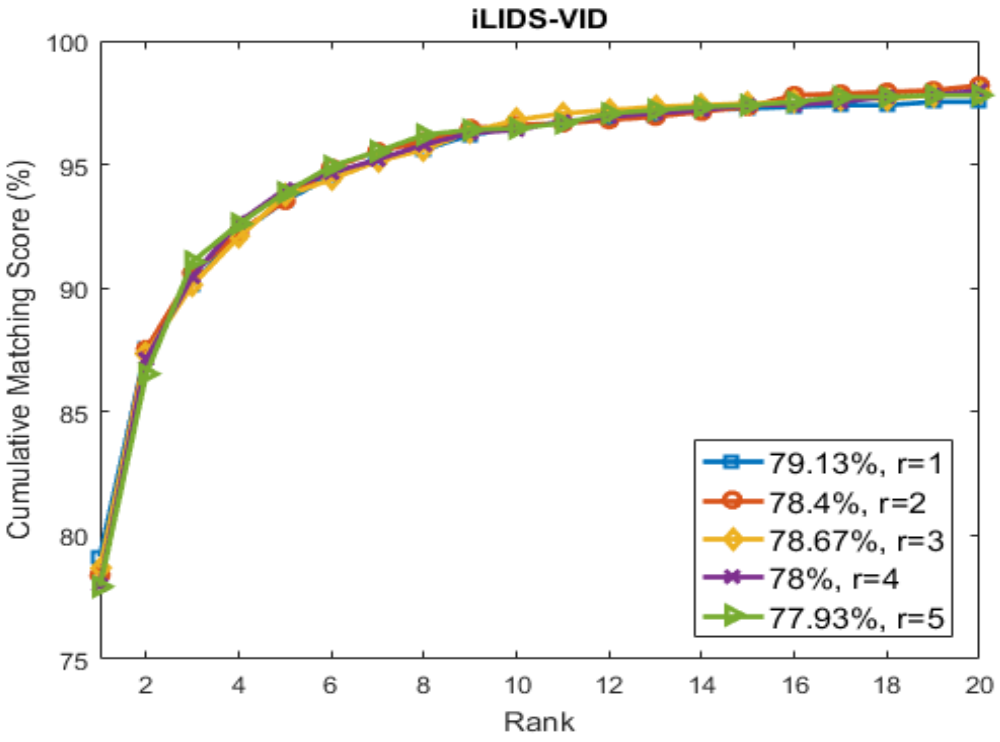


(a)

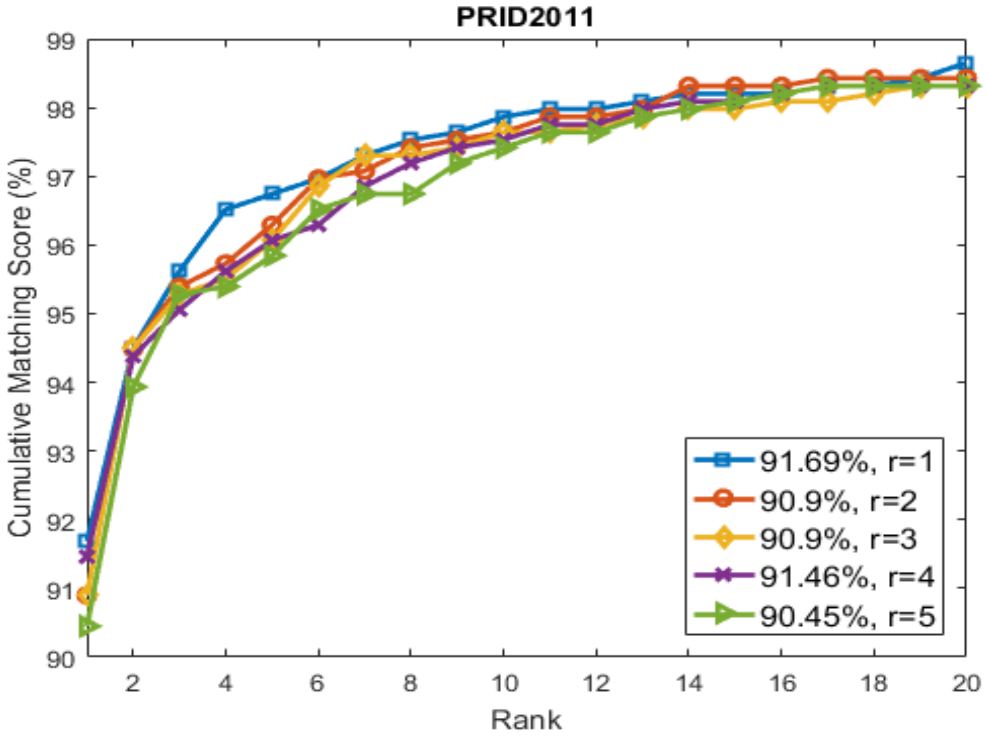


(b)

Figure 5.8: CMC curves comparing set-based matching vs. feature average-pooling on iLIDS-VID and PRID2011 datasets.



(a)



(b)

Figure 5.9: CMC curves obtained by varying the number of nearest neighbours from 1 to 5 on iLIDS-VID and PRID2011 datasets.

Figure 5.8. The improvement the proposed matching method brings over average-pooling is more consistent on the PRID2011 dataset although it is observed on both datasets for rank-1 accuracy. Since PRID2011 includes less challenges than iLIDS-VID, deriving meaningful clusters on the former is easier as these are more likely to reflect a specific attribute. In other words, it is more difficult for k-means algorithm to attain semantically meaningful clusters when many challenging attributes are combined together.

Number of Nearest Neighbours

It has been argued previously that the number of nearest neighbours considered in the NBNN algorithm barely affects the system's performance [199]. This was also validated experimentally as can be seen in Figure 5.9 where the number of nearest neighbours (parameter r) was varied between 1 and 5. Small difference or slightly worse performance is observed for a bigger r, which encourages the use of $r = 1$ to avoid unnecessary added complexity.

5.4.4 Computational Cost

The proposed system was implemented in MATLAB on a desktop PC with Intel Xeon CPU @ 3.60GHz and 64GB RAM. In the following, the matching time is reported similarly to previous related works [155, 179]. For the testing phase, the time taken to compute the similarity between two tracklets is approximately 0.015 seconds. Therefore, the cost of finding the match for a given probe depends on the size of the gallery set. For instance, for PRID2011 dataset, it takes almost 1.35 seconds to generate a ranking list of the gallery elements with respect to a given probe. This indicates that the proposed method is in fact efficient.

5.5 Summary

This chapter described a novel method for unsupervised video-based person re-id. By regarding the task as a classification problem, a set-based distance measure that can better cope with noisy uninformative data was developed. More importantly, the superiority of Spearman distance was demonstrated when used in conjunction with GOG3D features. Spearman distance is a rank-based distance measure that is almost completely ignored in the literature compared to other distance measures, particularly the Euclidean distance. An evaluation of the proposed method on

three public benchmarks was conducted achieving outstanding results on two challenging small datasets, and competitive results on the large-scale benchmark. Different components of the proposed system were also thoroughly analysed highlighting the amount each of them contributes to the performance improvement.

Chapter 6

Conclusions

6.1 Introduction

A core component of an automated surveillance system, person re-id, was addressed in this thesis. Despite recent progress in terms of accuracy, the algorithms concerned are mostly supervised deep learning methods that require a substantial amount of pairwise annotated data for training. Given the high cost and difficulty of obtaining such data from relative pairwise camera views, this requirement prevents real-world deployment of existing high-performing models. Therefore, the aim of this thesis was to bridge the gap between conventional re-id and real-world requirements. Several phases of the research led into achieving this goal. Firstly, forming a good understanding of the challenges associated with the problem was deemed necessary. For this purpose, a systematic evaluation was conducted highlighting the pertinent reasons affecting existing systems' accuracy. Insights from this work enabled the development of novel supervised and unsupervised methods for video-based person re-id. As the supervised model proved highly accurate on small annotated datasets, the unsupervised algorithm contributed significantly into bridging the performance gap with fully supervised models. Moreover, it is applicable off-the-shelf to any pair of camera views as no prior training is needed. Despite the advantages presented, additional room for improvement still exists. The accuracy is to be further improved especially on large-scale datasets of automatically detected and tracked persons. This could be achieved by providing a solution to the misalignment problem caused by automatic detection and tracking. Furthermore, integrating the re-id model into an end-to-end surveillance system would be critical to achieving

practical deployment.

6.2 Thesis Contributions

The development of this thesis led to the following contributions:

- An annotation strategy for challenging dataset attributes that may increase a person's cross-view appearance variation and contribute to re-id accuracy deterioration was developed. Six attributes were considered for this purpose: viewpoint angle variations, illumination changes, occlusions, background clutter, motion blur, and pose variations. Although the effect of these attributes on the accuracy is widely acknowledged within the re-id community [3, 23, 53, 161], this is the first work that attempts to quantify the impact each of them individually inflicts on the re-id performance. Following this work, an extensive evaluation was conducted in [161] to analyse the effect of varying viewpoints on the performance. As the authors of [161] expressed their willingness to extend their evaluation to other challenging attributes, gaining better insights on the inherent problems of the task offer solid grounds to resolve the remaining problems.

- A systematic evaluation of the most widely used image-based hand-crafted feature representations on a collection of four fully annotated datasets was conducted. The aim of this work was two-fold: (i) to identify the best performing image-based person descriptor with respect to each of the six challenging cases, and (ii) to unveil the persistent problems with existing feature representations that reduce their ability to deal with specific data attributes. This work would pave the way towards exploiting the advancements achieved in image-based re-id feature design to better tackle feature engineering in video-based re-id.

- A robust spatio-temporal person descriptor, i.e. feature extraction method for video-based person re-id, was proposed. Motivated by the findings of the preceding evaluation, the best performing image descriptor GOG [12] was improved and extended to three dimensions, thus obtaining GOG3D. In addition to colour and texture information that was shown to be insufficiently robust to cross-camera appearance variations, the proposed descriptor leveraged successfully the temporal information provided by video sequences for further discriminability between persons. The improvement achieved compared to similar spatio-temporal

descriptors was shown to be substantial.

- A supervised video re-id technique combining GOG3D with distance metric learning was investigated. The latter was the dominant direction in person re-id before the release of large-scale annotated datasets that triggered an abundance of deep learning techniques to take over. The Small Sample Size (SSS) problem was previously successfully tackled by the kNFST [108] algorithm for image-based re-id. We showed that when GOG3D is combined with kNFST, a simple and efficient algorithm that outperforms complex methods including deep learning on two challenging video datasets is obtained.

- For improved practical deployment, an unsupervised video person re-id method was developed. This method aims at matching cross-view person tracklets through a carefully designed set-based distance measure without the use of any identity labels. After clustering frame-wise feature vectors, the proposed method allows the rejection of outlier clusters for more adequate matching. Moreover, it is the first attempt to leverage a distance measure based on rank vectors such as the Spearman distance in a re-id application. In terms of performance, this method outperforms existing unsupervised methods, and closes the gap with supervised learning on two small datasets. It also competes with deep learning methods on a large-scale dataset.

6.3 Future Work

The person re-identification field has evolved to a great extent within the past few years. For instance, the highest reported rank-1 accuracy on the PRID2011 dataset upon the start of this project in 2016 was almost 64% [34], it has currently reached 94% [36]. Despite the problem nearing a solution in terms of accuracy, a number of factors still prevent real-world deployment. We identify the following promising directions for future work towards a more realistic application.

- GOG3D feature proposed in Chapter 4 is currently the best-performing hand-crafted spatio-temporal descriptor employed in person re-id. Nevertheless, it suffers from the misalignment problem caused by automatic detection and tracking of pedestrians as noticed when evaluated on the MARS dataset in Chapter 5. This is caused by the use of a rigid part model where body parts are pre-defined using equal-sized overlapping horizontal stripes for com-

putational convenience. For future work, it would be beneficial to take advantage of the progress achieved in human pose estimation [57, 58] in order to provide a better estimation of body parts from which GOG3D features can subsequently be extracted and fused.

- The dominance of deep learning over almost every aspect of computer vision lately shows a great potential that could equally be exploited in resolving the re-id problem. Given the current abundance of large-scale annotated re-id datasets, a promising future direction would be to tackle the weak generalisation ability of current systems to other domains, such as different camera views, cities, scenes, weather conditions, and time of the day, through deep transfer learning or unsupervised domain adaptation. These methods aim to learn a model on a source annotated dataset that can generalise well on a target unlabelled dataset. More research is being devoted towards this direction lately [136, 208, 209] but a lot of improvement is still needed to match the performance of fully supervised methods.

- In some forensic applications, there are cases where obtaining a single image of the person of interest is impossible. Alternatively, a description might be available from an eye-witness or a victim for instance, such as *tall male with dark hair and light skin wearing blue jeans and a red t-shirt*. Given such description, i.e. a set of semantic attributes, a search for the corresponding subject is undertaken in various camera networks. This is known as zero-shot re-identification. This problem was initially addressed in [210] where an ontology of 21 semantic attributes was introduced involving clothes colours, style and patterns, in addition to carried objects, hair and gender. However, when not fused with low-level colour and texture features, the performance produced was very poor. This was mainly caused by false attribute detections due to the poor quality of surveillance data. To enhance attribute detection for zero-shot re-id, a promising approach would be to leverage generative models such as Generative Adversarial Networks (GANs) in order to enhance the quality of the data. This could be followed by body part detection on which relevant attributes can be detected separately to reduce the errors.

- Instead of breaking down person re-id into separate task, a few attempts were made to date to integrate person detection and re-identification in an end-to-end system using deep CNNs [211, 212]. For a practical surveillance application, we would like to improve these models and integrate an action recognition module into this framework. A person can thus be

detected and tracked across a number of cameras while simultaneously identifying their actions, i.e. walking, running, sitting, etc. In addition to the importance of such a system in automating video surveillance, it is extremely useful for sports analysis applications where labelling the actions performed by individual players in real-time is extremely tedious and costly. Providing automatic tagging for sports games on-the-fly enables the flow of a sufficient amount of annotated data for better insights into individual players' performance.

References

[1] G. J. Smith, "Behind the screens: Examining constructions of deviance and informal practices among cctv control room operators in the uk," *Surveillance & Society*, vol. 2, no. 2/3, 2004.

[2] V. Tsakanikas and T. Dagiuklas, "Video surveillance systems-current status and future trends," *Computers & Electrical Engineering*, vol. 70, pp. 736–753, 2018.

[3] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *arXiv preprint arXiv:1610.02984*, 2016.

[4] J. Zhang, Y. Yuan, and Q. Wang, "Night person re-identification and a benchmark," *IEEE Access*, vol. 7, pp. 95496–95504, 2019.

[5] S. Gong, M. Cristani, S. Yan, and C. C. Loy, *Person Re-Identification*. Springer, 2014.

[6] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[7] D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in *European conference on computer vision (ECCV)*, 2008.

[8] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2360–2367, IEEE, 2010.

[9] B. Ma, Y. Su, and F. Jurie, "Bicov: a novel image representation for person re-identification and face verification," in *British Machive Vision Conference (BMVC)*, 2012.

[10] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[11] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] T. Matsukawa, T. Okabe, E. Suzuki, and Y. Sato, "Hierarchical gaussian descriptor for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] S. Karanam, Y. Li, and R. J. Radke, "Sparse re-id: Block sparsity for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015.

[14] S. Karanam, Z. Wu, and R. J. Radke, "Learning affine hull representations for multi-shot person re-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.

[15] X. Zhu, X.-Y. Jing, F. Wu, Y. Wang, W. Zuo, and W.-S. Zheng, "Learning heterogeneous dictionary pair with feature projection matrix for pedestrian video retrieval via single query image," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[16] T. Li, L. Sun, C. Han, and J. Guo, "Salient region-based least-squares log-density gradient clustering for image-to-video person re-identification," *IEEE Access*, vol. 6, pp. 8638–8648, 2018.

[17] D. Zhang, W. Wu, H. Cheng, R. Zhang, Z. Dong, and Z. Cai, "Image-to-video person re-identification with temporally memorized similarity learning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2622–2632, 2017.

[18] B. Yu, N. Xu, and J. Zhou, "Cross-media body-part attention network for image-to-video person re-identification," *IEEE Access*, vol. 7, pp. 94966–94976, 2019.

[19] G. Wang, J. Lai, and X. Xie, "P2snet: can an image match a video for person re-identification in an end-to-end way?," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2777–2787, 2017.

[20] B. Cancela, T. M. Hospedales, and S. Gong, "Open-world person re-identification by multi-label assignment inference," in *British Machine Vision Conference (BMVC)*, 2014.

[21] W.-S. Zheng, S. Gong, and T. Xiang, "Towards open-world person re-identification by one-shot group-based verification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 591–606, 2015.

[22] X. Li, A. Wu, and W.-S. Zheng, "Adversarial open-world person re-identification," in *European Conference on Computer Vision (ECCV)*, 2018.

[23] S. Gong, M. Cristani, C. C. Loy, and T. M. Hospedales, "The re-identification challenge," in *Person re-identification*, pp. 1–20, Springer, 2014.

[24] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *International Conference on Computer Vision (ICCV)*, 2015.

[25] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[26] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-s. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[27] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[28] S. Bai, P. Tang, P. H. Torr, and L. J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 740–749, 2019.

[29] Y. Sun, Q. Xu, Y. Li, C. Zhang, Y. Li, S. Wang, and J. Sun, "Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Vrstc: Occlusion-free video

person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[31] D. Chen, H. Li, T. Xiao, S. Yi, and X. Wang, "Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[32] R. Hou, B. Ma, H. Chang, X. Gu, S. Shan, and X. Chen, "Interaction-and-aggregation network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[33] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by video ranking," in *European Conference on Computer Vision (ECCV)*, 2014.

[34] K. Liu, B. Ma, W. Zhang, and R. Huang, "A spatio-temporal appearance representation for video-based pedestrian re-identification," in *International Conference on Computer Vision (ICCV)*, 2015.

[35] C. Riachy and A. Bouridane, "Person re-identification: Attribute-based feature evaluation," in *IEEE World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2018.

[36] C. Riachy, N. Al-Maadeed, D. Organisciak, F. Khelifi, and B. Ahmed, "3d gaussian descriptor for video-based person re-identification," in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2019.

[37] C. Riachy, F. Khelifi, and A. Bouridane, "Video-based person re-identification using unsupervised tracklet matching," *IEEE Access*, vol. 7, pp. 20596–20606, 2019.

[38] D. J. Felleman and D. E. Van, "Distributed hierarchical processing in the primate cerebral cortex.," *Cerebral cortex (New York, NY: 1991)*, vol. 1, no. 1, pp. 1–47, 1991.

[39] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron*, vol. 73, no. 3, pp. 415–434, 2012.

[40] N. Gheissari, T. B. Sebastian, and R. Hartley, "Person re-identification using spatiotemporal appearance," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2006.

[41] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognition Letters*, vol. 34, no. 1, pp. 3–19, 2013.

[42] O. Javed, K. Shafique, Z. Rasheed, and M. Shah, "Modeling inter-camera space–time and appearance relationships for tracking across non-overlapping views," *Computer Vision and Image Understanding*, vol. 109, no. 2, pp. 146–162, 2008.

[43] R. Mazzon, S. F. Tahir, and A. Cavallaro, "Person re-identification in crowd," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1828–1837, 2012.

[44] C. C. Loy, T. Xiang, and S. Gong, "Multi-camera activity correlation analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1988–1995, IEEE, 2009.

[45] D. Makris, T. Ellis, and J. Black, "Bridging the gaps between cameras," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2004.

[46] A. Rahimi, B. Dunagan, and T. Darrell, "Simultaneous calibration and tracking with a network of non-overlapping sensors," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2004.

[47] M. Lantagne, M. Parizeau, and R. Bergevin, "Vip: Vision tool for comparing images of people," in *Vision Interface*, vol. 2, 2003.

[48] U. Park, A. K. Jain, I. Kitahara, K. Kogure, and N. Hagita, "Vise: Visual search engine using multiple networked cameras," in *18th International Conference on Pattern Recognition (ICPR)*, IEEE, 2006.

[49] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 663–671, 2006.

[50] F. Porikli, "Inter-camera color calibration by correlation model function," in *Proceedings 2003 International Conference on Image Processing (ICIP)*, IEEE, 2003.

[51] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof, "Person re-identification by descriptive and discriminative classification," in *Scandinavian Conference on Image Analysis (SCIA)*, 2011.

[52] W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[53] S. Karanam, M. Gou, Z. Wu, A. Rates-Borras, O. Camps, and R. J. Radke, "A systematic evaluation and benchmark for person re-identification: Features, metrics, and datasets," *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, 2019.

[54] F. Xiong, M. Gou, O. Camps, and M. Sznaier, "Person re-identification using kernel-based metric learning methods," in *European Conference on Computer Vision (ECCV)*, 2014.

[55] A. Mignon and F. Jurie, "Pcca: A new approach for distance learning from sparse pairwise constraints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[56] A. Das, A. Chakraborty, and A. K. Roy-Chowdhury, "Consistent re-identification in a camera network," in *European Conference on Computer Vision (ECCV)*, Springer, 2014.

[57] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[58] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[59] R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 144–151, 2014.

[60] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, and S. Z. Li, "Salient color names for person re-identification," in *European conference on computer vision (ECCV)*, Springer, 2014.

[61] G. Lisanti, I. Masi, A. D. Bagdanov, and A. Del Bimbo, "Person re-identification by iterative re-weighted sparse ranking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 8, pp. 1629–1642, 2014.

[62] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[63] R. Zhao, W. Ouyang, and X. Wang, "Person re-identification by salience matching," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2528–2535, 2013.

[64] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 39, no. 2, pp. 356–370, 2017.

[65] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision (IJCV)*, vol. 60, no. 2, pp. 91–110, 2004.

[66] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 971–987, 2002.

[67] J. Van De Weijer, C. Schmid, J. Verbeek, and D. Larlus, "Learning color names for real-world applications," *IEEE Transactions on Image Processing (TIP)*, vol. 18, no. 7, pp. 1512–1523, 2009.

[68] B. Ma, Y. Su, and F. Jurie, "Covariance descriptor based on bio-inspired features for person re-identification and face verification," *Image and Vision Computing*, vol. 32, no. 6-7, pp. 379–390, 2014.

[69] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, p. 1019, 1999.

[70] S. Liao, G. Zhao, V. Kellokumpu, M. Pietikäinen, and S. Z. Li, "Modeling pixel process with scale invariant local patterns for background subtraction in complex scenes," in *Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1301–1306, IEEE, 2010.

[71] S. Pedagadi, J. Orwell, S. Velastin, and B. Boghossian, "Local fisher discriminant analysis for pedestrian re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[72] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, and J. R. Smith, "Learning locally-adaptive decision functions for person verification," in *IEEE Conference on Computer Vision and Pattern Recognition (VPR)*, 2013.

[73] X. Liu, M. Song, D. Tao, X. Zhou, C. Chen, and J. Bu, "Semi-supervised coupled dictionary learning for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[74] M. Gou, O. Camps, and M. Sznaier, "mom: Mean of moments feature for person re-identification," in *International Conference on Computer Vision (ICCV)*, 2017.

[75] R. Layne, T. M. Hospedales, and S. Gong, "Person re-identification by attributes.," in *British Machine Vision Conference (BMVC)*, 2012.

[76] R. Layne, T. M. Hospedales, and S. Gong, "Re-id: Hunting attributes in the wild," in *British Machine Vision Conference (BMVC)*, 2014.

[77] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao, "Multi-task learning with low rank attribute embedding for person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[78] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Z. Hu, C. Yan, and Y. Yang, "Improving person re-identification by attribute and identity learning," *Pattern Recognition*, 2019.

[79] A. Schumann and R. Stiefelhagen, "Person re-identification by deep learning attribute-complementary information," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017.

[80] J. Wang, X. Zhu, S. Gong, and W. Li, "Transferable joint attribute-identity deep learning for unsupervised person re-identification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[81] Z. Yin, W.-S. Zheng, A. Wu, H.-X. Yu, H. Wan, X. Guo, F. Huang, and J. Lai, "Adversarial attribute-image person re-identification," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1100–1106, 2018.

[82] Y. Zhang, X. Gu, J. Tang, K. Cheng, and S. Tan, "Part-based attribute-aware network for person re-identification," *IEEE Access*, vol. 7, pp. 53585–53595, 2019.

[83] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," in *International conference on database theory*, pp. 217–235, Springer, 1999.

[84] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional space," in *International conference on database theory*, pp. 420–434, 2001.

[85] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[86] Y. Yang, S. Liao, Z. Lei, and S. Z. Li, "Large scale similarity learning using similar pairs for person verification," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[87] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of educational psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[88] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, 2006.

[89] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, "Distance metric learning with application to clustering with side-information," in *Advances in neural information processing systems*, pp. 521–528, 2003.

[90] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *international conference on Machine learning (ICML)*, ACM, 2004.

[91] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *international conference on Machine learning (ICML)*, ACM, 2007.

[92] S. Xiang, F. Nie, and C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern recognition*, vol. 41, no. 12, pp. 3600–3612, 2008.

[93] A. Globerson and S. T. Roweis, "Metric learning by collapsing classes," in *Advances in neural information processing systems*, pp. 451–458, 2006.

[94] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[95] B. Kulis, P. Jain, and K. Grauman, "Fast similarity search for learned metrics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 12, pp. 2143–2157, 2009.

[96] S. C. Hoi, W. Liu, M. R. Lyu, and W.-Y. Ma, "Learning distance metrics with contextual constraints for image retrieval," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2006.

[97] S. C. Hoi, W. Liu, and S.-F. Chang, "Semi-supervised distance metric learning for collaborative image retrieval and clustering," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 6, no. 3, 2010.

[98] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *international conference on computer vision (ICCV)*, pp. 498–505, IEEE, 2009.

[99] H. V. Nguyen and L. Bai, "Cosine similarity metric learning for face verification," in *Asian conference on computer vision (ACCV)*, Springer, 2010.

[100] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[101] W.-S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPr)*, 2011.

[102] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof, "Relaxed pairwise learned metric for person re-identification," in *European conference on computer vision (ECCV)*, 2012.

[103] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[104] M. Sugiyama, "Local fisher discriminant analysis for supervised dimensionality reduction," in *international conference on Machine learning (ICML)*, 2006.

[105] X. He and P. Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, pp. 153–160, 2004.

[106] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and exten-

sions: A general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 1, pp. 40–51, 2007.

[107] Y.-F. Guo, L. Wu, H. Lu, Z. Feng, and X. Xue, "Null foley–sammon transform," *Pattern recognition*, vol. 39, no. 11, pp. 2248–2251, 2006.

[108] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[109] B. J. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary, "Person re-identification by support vector ranking.," in *British Machine Vision Conference (BMVC)*, 2010.

[110] Y. Zhang, B. Li, H. Lu, A. Irie, and X. Ruan, "Sample-specific svm learning for person re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1278–1287, 2016.

[111] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational intelligence and neuroscience*, 2018.

[112] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3908–3916, 2015.

[113] R. R. Varior, M. Haloi, and G. Wang, "Gated siamese convolutional neural network architecture for human re-identification," in *European conference on computer vision (ECCV)*, 2016.

[114] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based cnn with improved triplet loss function," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[115] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, and X. Tang, "Spindle net: Person re-identification with human body region guided feature decomposition and fusion," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[116] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[117] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *European Conference on Computer Vision (ECCV)*, 2018.

[118] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, "Performance measures and a data set for multi-target, multi-camera tracking," in *European Conference on Computer Vision (ECCV)*, 2016.

[119] M. Zheng, S. Karanam, Z. Wu, and R. J. Radke, "Re-identification with consistent attentive siamese networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[120] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[121] J. Liu, Z.-J. Zha, Q. Tian, D. Liu, T. Yao, Q. Ling, and T. Mei, "Multi-scale triplet cnn for person re-identification," in *Proceedings of the 24th ACM international conference on Multimedia*, pp. 192–196, ACM, 2016.

[122] C. Su, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Deep attributes driven multi-camera person re-identification," in *European conference on computer vision (ECCV)*, 2016.

[123] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[124] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[125] F. Wang, W. Zuo, L. Lin, D. Zhang, and L. Zhang, "Joint learning of single-image and cross-image representations for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[126] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[127] D. Organisciak, C. Riachy, N. Aslam, and H. Shum, "Triplet loss with channel attention for person re-identification," in *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, 2019.

[128] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[129] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[130] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, and S. Z. Li, "Embedding deep metric for person re-identification: A study against large variations," in *European conference on computer vision (ECCV)*, 2016.

[131] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[132] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[133] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[134] W. Deng, L. Zheng, Q. Ye, G. Kang, Y. Yang, and J. Jiao, "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in *EEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[135] S. Bak, P. Carr, and J.-F. Lalonde, "Domain adaptation through synthesis for unsupervised person re-identification," in *European Conference on Computer Vision (ECCV)*, 2018.

[136] J. Liu, Z.-J. Zha, D. Chen, R. Hong, and M. Wang, "Adaptive transfer network for cross-domain person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[137] Y.-J. Cho and K.-J. Yoon, "Improving person re-identification via pose-aware multi-shot matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[138] W. Huang, C. Liang, Y. Yu, Z. Wang, W. Ruan, and R. Hu, "Video-based person re-identification via self paced weighting," in *AAAI Conference on Artificial Intelligence*, 2018.

[139] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[140] Y. Liu, J. Yan, and W. Ouyang, "Quality aware network for set to set recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[141] N. McLaughlin, J. Martinez del Rincon, and P. Miller, "Recurrent convolutional network for video-based person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[142] S. Xu, Y. Cheng, K. Gu, Y. Yang, S. Chang, and P. Zhou, "Jointly attentive spatial-temporal pooling networks for video-based person re-identification," in *International Conference on Computer Vision (ICCV)*, 2017.

[143] J. You, A. Wu, X. Li, and W.-S. Zheng, "Top-push video-based person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[144] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[145] F. M. Khan and F. Brèmond, "Multi-shot person re-identification using part appearance mixture," in *Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[146] S. Karanam, Y. Li, and R. J. Radke, "Person re-identification with discriminatively trained viewpoint invariant dictionaries," in *International Conference on Computer Vision (ICCV)*, 2015.

[147] Y. Chen, X. Zhu, and S. Gong, "Deep association learning for unsupervised video person re-identification," in *British Machine Vision Conference (BMVC)*, 2018.

[148] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, "Mars: A video bench-

mark for large-scale person re-identification," in *European Conference on Computer Vision (ECCV)*, 2016.

[149] T. Wang, S. Gong, X. Zhu, and S. Wang, "Person re-identification by discriminative selection in video ranking," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 12, pp. 2501–2514, 2016.

[150] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference (BMVC)*, 2008.

[151] Z. Liu, J. Chen, and Y. Wang, "A fast adaptive spatio-temporal 3d feature for video-based person re-identification," in *IEEE International Conference on Image Processing (ICIP)*, 2016.

[152] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, 1967.

[153] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," tech. rep., Stanford, 2006.

[154] X. Zhu, X.-Y. Jing, X. You, X. Zhang, and T. Zhang, "Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5683–5695, 2018.

[155] Z. Liu, D. Wang, and H. Lu, "Stepwise metric promotion for unsupervised video person re-identification," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[156] Y. Suh, J. Wang, S. Tang, T. Mei, and K. M. Lee, "Part-aligned bilinear representations for person re-identification," in *European Conference on Computer Vision (ECCV)*, 2018.

[157] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[158] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[159] J. Si, H. Zhang, C.-G. Li, J. Kuen, X. Kong, A. C. Kot, and G. Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[160] K. Kishida, *Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments*. National Institute of Informatics Tokyo, Japan, 2005.

[161] X. Sun and L. Zheng, "Dissecting person re-identification from the viewpoint of viewpoint," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[162] D. Gabor, "Theory of communication. part 1: The analysis of information," *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, vol. 93, no. 26, pp. 429–441, 1946.

[163] C. Schmid, "Constructing models for content-based image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. II–II, IEEE, 2001.

[164] C. Adak, "Gabor filter and rough clustering based edge detection," in *International Conference on Human Computer Interactions (ICHCI)*, 2013.

[165] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 93–104, 2008.

[166] G. Guo, G. Mu, Y. Fu, and T. S. Huang, "Human age estimation using bio-inspired features," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 112–119, 2009.

[167] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 2037–2041, 2006.

[168] C.-H. Chan, J. Kittler, and K. Messer, "Multi-scale local binary pattern histograms for face recognition," in *International conference on biometrics*, pp. 809–818, Springer, 2007.

[169] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[170] S. Liao, W. Fan, A. C. Chung, and D.-Y. Yeung, "Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features," in *International Conference on Image Processing (ICIP)*, pp. 665–668, IEEE, 2006.

[171] X. Wang, T. X. Han, and S. Yan, "An hog-lbp human detector with partial occlusion handling," in *International Conference on Computer Vision (ICCV)*, pp. 32–39, IEEE, 2009.

[172] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 915–928, 2007.

[173] M. Bicego, A. Lagorio, E. Grosso, and M. Tistarelli, "On the use of sift features for face authentication," in *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pp. 35–35, IEEE, 2006.

[174] U. Park, S. Pankanti, and A. K. Jain, "Fingerprint verification using sift features," in *Biometric Technology for Human Identification V*, vol. 6944, p. 69440K, International Society for Optics and Photonics, 2008.

[175] F. Alonso-Fernandez, P. Tome-Gonzalez, V. Ruiz-Albacete, and J. Ortega-Garcia, "Iris recognition based on sift features," in *2009 First IEEE International Conference on Biometrics, Identity and Security (BIdS)*, pp. 1–8, IEEE, 2009.

[176] D. J. Jobson, Z.-u. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Transactions on Image processing*, vol. 6, no. 7, pp. 965–976, 1997.

[177] B. Bascle, O. Bernier, and V. Lemaire, "Illumination-invariant color image correction," in *International Workshop on Intelligent Computing in Pattern Analysis and Synthesis*, pp. 359–368, Springer, 2006.

[178] P. M. Roth, M. Hirzer, M. Köstinger, C. Beleznai, and H. Bischof, "Mahalanobis distance learning for person re-identification," in *Person re-identification*, pp. 247–267, Springer, 2014.

[179] M. Ye, A. J. Ma, L. Zheng, J. Li, and P. C. Yuen, "Dynamic label graph matching for unsu-

pervised video re-identification," in *International Conference on Computer Vision (ICCV)*, 2017.

[180] M. Li, X. Zhu, and S. Gong, "Unsupervised person re-identification by deep learning tracklet association," in *European Conference on Computer Vision (ECCV)*, 2018.

[181] Z. Zhou, Y. Huang, W. Wang, L. Wang, and T. Tan, "See the forest for the trees: Joint spatial and temporal recurrent neural networks for video-based person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[182] J. Xu, R. Zhao, F. Zhu, H. Wang, and W. Ouyang, "Attention-aware compositional network for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[183] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[184] X. Chang, T. M. Hospedales, and T. Xiang, "Multi-level factorisation net for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[185] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, pp. 357–360, ACM, 2007.

[186] B. Ma, Q. Li, and H. Chang, "Gaussian descriptor based on local features for person re-identification," in *Asian Conference on Computer Vision (ACCV)*, 2014.

[187] T. Kobayashi and N. Otsu, "Image feature extraction using gradient local auto-correlations," in *European Conference on Computer Vision (ECCV)*, 2008.

[188] J. N. Kather, A. Weidner, U. Attenberger, Y. Bukschat, C.-A. Weis, M. Weis, L. R. Schad, and F. G. Zöllner, "Color-coded visualization of magnetic resonance imaging multiparametric maps," *Scientific reports*, vol. 7, p. 41107, 2017.

[189] L. Bazzani, M. Cristani, and V. Murino, "Symmetry-driven accumulation of local features for human characterization and re-identification," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 130–144, 2013.

[190] M. Lovrić, M. Min-Oo, and E. A. Ruh, "Multivariate normal distributions parametrized as

a riemannian symmetric space," *Journal of Multivariate Analysis*, vol. 74, no. 1, pp. 36–48, 2000.

[191] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision (ECCV)*, 2010.

[192] H. Jégou and O. Chum, "Negative evidences and co-occurences in image retrieval: The benefit of pca and whitening," in *European conference on computer vision (ECCV)*, 2012.

[193] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[194] L. Xie, Q. Tian, and B. Zhang, "Simple techniques make sense: Feature pooling and normalization for image classification," *IEEE transactions on circuits and systems for video technology*, vol. 26, no. 7, pp. 1251–1264, 2016.

[195] E. Hoffer, R. Banner, I. Golan, and D. Soudry, "Norm matters: efficient and accurate normalization schemes in deep networks," in *Advances in Neural Information Processing Systems*, 2018.

[196] F. M. Khan and F. Bremond, "Unsupervised data association for metric learning in the context of multi-shot person re-identification," in *Advanced Video and Signal Based Surveillance (AVSS)*, 2016.

[197] Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep kronecker-product matching for person re-identification," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[198] M. Ye, X. Lan, and P. C. Yuen, "Robust anchor embedding for unsupervised video person re-identification in the wild," in *European Conference on Computer Vision (ECCV)*, 2018.

[199] O. Boiman, E. Shechtman, and M. Irani, "In defense of nearest-neighbor based image classification," in *Computer Vision and Pattern Recognition (CVPR)*, 2008.

[200] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, and Y. Zhong, "Person re-identification by unsupervised video matching," *Pattern Recognition*, vol. 65, pp. 197–210, 2017.

[201] S. McCann and D. G. Lowe, "Local naive bayes nearest neighbor for image classification," in *Computer Vision and Pattern Recognition (CVPR)*, 2012.

[202] J. Weng, C. Weng, and J. Yuan, "Spatio-temporal naive-bayes nearest-neighbor (st-nbnn) for skeleton-based action recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[203] X. Yang and Y. L. Tian, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2012.

[204] M. Ye, C. Liang, Y. Yu, Z. Wang, Q. Leng, C. Xiao, J. Chen, and R. Hu, "Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2553–2566, 2016.

[205] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Computer Vision and Pattern Recognition (CVPR)*, 2017.

[206] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, vol. 32, no. 9, pp. 1627–1645, 2010.

[207] A. Dehghan, S. Modiri Assari, and M. Shah, "Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking," in *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[208] J. Jia, Q. Ruan, and T. M. Hospedales, "Frustratingly easy person re-identification: Generalizing person re-id in practice," *British Machine Vision Conference (BMVC)*, 2019.

[209] J. Song, Y. Yang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Generalizable person re-identification by domain-invariant mapping network," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[210] R. Layne, T. M. Hospedales, and S. Gong, "Attributes-based re-identification," in *Person Re-Identification*, pp. 93–117, Springer, 2014.

[211] Z. He and L. Zhang, "End-to-end detection and re-identification integrated net for person search," in *Asian Conference on Computer Vision (ACCV)*, 2018.

[212] B. Munjal, S. Amin, F. Tombari, and F. Galasso, "Query-guided end-to-end person search," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.